The Confidence-Accuracy Relationship Using Scale Versus Other Methods of Assessing Confidence

Jamal K. Mansour

Author Note

Jamal K. Mansour, Memory Research Group, Centre for Applied Social Sciences; Psychology, Sociology, and Education; Queen Margaret University.

This research was funded by two grants to the author from the Centre for Applied Social Sciences at Queen Margaret University.

Correspondence concerning this article should be addressed to Jamal K. Mansour, Psychology, Sociology, & Education, Queen Margaret University, Edinburgh, UK EH21 6UU Phone: +44 (0)131 474 0 000. Fax: +44 (0)131 474 0001. Email: jmansour@qmu.ac.uk

Abstract

Historically, researchers have collected eyewitness identification confidence using scales; however, in practice, eyewitnesses are more commonly asked for a verbal statement. In Experiment 1, participants viewed a simultaneous lineup and provided confidence in their own words, by explaining why they made their decision, or by selecting from statements made by real eyewitnesses, and then provided a scale rating (0-100%) or provided only the scale rating. In Experiment 2, participants viewed a sequential lineup and provided confidence in their own words followed by the scale rating or only the scale rating. Confidence predicted identification accuracy in all conditions, although verbal statements were highly variable and challenging to interpret. For example, only when scale-based confidence was high (80%+) did interpretation of the verbal confidence statement reliably align with scale-based confidence. These data highlight the complexity of verbal confidence statements and the need to establish meaningful boundaries for interpreting verbal confidence statements.

Keywords: confidence, eyewitness identification, CAC, calibration

General Audience Summary

Researchers commonly ask participants who view a lineup in an experiment to indicate how confident they are in their decision using a scale (e.g., 0-100%; Not at all confident-Completely confident). However, in practice, eyewitnesses may be asked to make a verbal statement—to give their confidence in their own words or to explain why they made their decision. Two experiments examined whether the relationship between confidence and accuracy was similar across different methods of asking about confidence. In Experiment 1, participants made a simultaneous lineup decision and then indicated their confidence in their own words, indicated why they made their decision, or selected a statement representative of their confidence from amongst statements made by real eyewitnesses, and then rated their confidence on a 0-100% scale. Other participants just rated their confidence on the scale. In Experiment 2, participants made a sequential lineup decision and then indicated their confidence in their own words followed by the scale or just using the scale. Confidence was predictive of accuracy regardless of the way it was collected. However, the data raise concerns about how to make use of verbal confidence statements as interpretations of these statements varied considerably and, particularly when confidence was low or medium, did not reliably align with the level of confidence that the eyewitness may have intended to convey. Meaningful boundaries for interpreting verbal statements of confidence are needed.

The Confidence-Accuracy Relationship Using Scale Versus Other Methods of Assessing Confidence

Triers of fact primarily rely on an eyewitness' confidence as an indicator of the accuracy of their identification (Cutler, Penrod, & Dexter, 1990; Cutler, Penrod, & Stuve, 1988; *Manson v. Braithwaite*, 1977). For many years, eyewitness identification confidence was understood to be, at best, moderately related to accuracy (Leippe & Eisenstadt, 2007). In the 1990s, it became apparent that the relationship was stronger for identifications than rejections or both combined (Sporer, 1993; Sporer, Penrod, Read, & Cutler, 1995); however, even for identifications the correlation was only moderate. Since then, researchers have demonstrated that the confidence-accuracy (CA) relationship for identifications can be stronger than suggested by the CA correlation (e.g., Juslin, Olsson, & Winman, 1996; Brewer & Wells, 2006). Adjusting calibration curves to use only suspect identifications—confidence-accuracy characteristic (CAC) curves—suggests the probative value of confidence is even greater (Mickes, 2015; Wixted, Mickes, Clark, Gronlund, & Roediger, 2015; Wixted & Wells, 2017). Moreover, Wixted and Wells demonstrated that across 15 studies on the CA relationship, high-confidence rejections were highly likely to be accurate.

Wixted and Wells (2017) argue that a strong CA relationship is present when the eyewitness' memory has not been contaminated prior to the confidence judgment. Indeed, feedback to eyewitnesses that they chose the suspect (Wells & Bradfield, 1998) and variability in lineup fairness (Sauer, Palmer, & Brewer, 2019) reduce the CA relationship. When an eyewitness' memory has not been contaminated, the CA relationship across participants can be strong despite differences in witnessing conditions (Carlson et al., 2016; Semmler, Dunn, Mickes, & Wixted, 2018; Wixted, Read, & Lindsay, 2016) but other times varies (Sauer, et al., 2019; Jalava, Smith, & Wells, under review; Smith, Wilford, Quigley-McBride, & Wells, 2019). We do not yet have a clear picture of the boundary conditions for

when confidence is a satisfactory predictor of lineup decision accuracy. Caution is also merited because, as Sauer et al. recently highlighted, high confidence indicates a *high probability* of accuracy, not *definite* accuracy.

There is a further challenge in generalizing findings about the CA relationship to practice. Most research on confidence has used numeric or verbal scales (e.g., 0-100%, Not at all confident-Completely confident). Although the CA relationship is unaffected by type of scale (Dodson & Dobolyi, 2016; Weber, Brewer, & Margitich, 2008), identification researchers have not compared scale and free report verbal statements. In practice, confidence is primarily collected verbally. In the United States, the standard and recommended practice is to request confidence in the eyewitness' own words (National Research Council, 2014). The police ask UK eyewitnesses (excluding Scotland) to confirm their identification (PACE Code D, 2017). In Scotland, eyewitnesses indicate why they identified the person they did.

At a glance, asking eyewitnesses why they made their identification appears to provide evidence supporting the eyewitness' decision, the strength of which could be evaluated by triers of fact. However, retrospective explanations of decision processes are frequently inferences based on the decision itself (Hall, Johansson, & Strandberg, 2012; Johansson, Hall, Sikström, Tärning, & Lind, 2006; Nisbett & Wilson, 1977). For example, Johansson et al. found participants rarely detected when their previously-made decision was modified and provided a reason for the modified decision when asked to do so. Ericsson and Simon (1980; 1993) proposed that when people describe or explain their cognitive processes retrospectively, they cannot remain focused on the task. Verbalization itself changes the thought processes that generated the original response. Thus, an explanation of a decision is based on a different sequence of thoughts than the decision itself (see also Ericsson, 2003, Figure 2).

However, confidence in (cf. explanation of) a decision can be predictive of accuracy

(Busey, Tunnicliff, Loftus, & Loftus, 2000; Juslin et al., 1996). Outside the eyewitness domain, verbal and numeric judgments of confidence have been shown to be similarly predictive, although numeric judgments are less variable (Budescu & Wallsten, 1990; Budescu, Weinberg, & Wallsten, 1988; Dhami & Wallsten, 2005; Karelitz & Budescu, 2004) and less affected by context and framing (Brun & Teigen, 1988; Windschitl & Wells, 1996). Furthermore, people use a large variety of probability phrases to express confidence. For example, Dhami and Wallsten (2005) obtained 102 distinct phrases from 29 participants asked to provide seven probability phrases each, indicating that people have quite different probability lexicons (see also, Wallsten & Budescu, 1995). Although people appear to have quite stable individual probability lexicons (Budescu, et al., 1988), a phrase's membership function—the probabilities covered by the phrase and the density at specific probabilities—is quite broad (Wallsten, Budescu, & Zwick, 1993; Wallsten, Budescu, Rapport, Zwick, & Forsyth, 1986). Furthermore, people do not appreciate the variability in interpretations of specific probability statements (Brun & Teigen, 1988; Murphy, Lichtenstein, Fischhoff, & Winkler, 1980).

Despite the advantages of numeric judgments, people—including eyewitnesses (Kenchel, Reisberg, & Dodson, 2017)—prefer to communicate confidence verbally (but prefer to receive confidence judgments numerically; Wallsten & Budescu, 1995). This preference reflects that people prefer to use vague terms when there is uncertainty in their judgement (e.g., Barnes, 2016; Wallsten & Budescu, 1995)—as may be expected when expressing confidence in an eyewitness identification. Encouragingly, the precision and reliability of probability estimates change little when using fine-grained (e.g., 0-100%) versus coarse-grained (e.g., 1-7) categories (Brun & Teigen, 1988; Dodson & Dobolyi, 2016). Commonly when verbal confidence judgements have been examined in the eyewitness literature, they have been categorized as low, medium, or high confidence (e.g., Wixted et al.,

2015; Wixted, Mickes, Dunn, Clark, & Wells, 2016). Thus, there is no reason to expect a difference in terms of how well numeric versus verbal confidence judgements predict identification accuracy; however, interpreting verbal confidence in lineup decisions is likely to be challenging.

This research examines the CA relationship using different ways of querying confidence—with the particular goal of comparing approaches used in practice to the approach used in research. I hypothesized that confidence and accuracy in identifications would be strongly related for scale ratings and verbal statements of confidence but weakly related (if at all) when participants were asked why they made their decision.[1] Moreover, I hypothesised that verbal statements of confidence in identifications would generally align with scale judgements but that providing reasons for identifying would less reliably align with scale judgments. I further hypothesized that scale ratings of confidence would be influenced by prior provision of a reason for the decision but not by being asked to provide a rating "in your own words" or selecting from a series of statements. Finally, I expected to replicate the finding that the CA relationship is higher for identifications than rejections but that highly confident rejections would be associated with a high probability of accuracy.

Experiment 1

Method

Participants. Amazon Mechanical Turk/Turk Prime was used to recruit participants. All workers had to have a HIT approval rate greater than 75% and have had more than 100 HITs approved. The usable sample (N = 968) did not include cases where the participants did not complete the experiment (n = 75), completed the experiment multiple times (based on IP addresses; (n = 57/75 cases), or did not consent to participate (n = 1). The usable sample of

[1] Hypotheses were preregistered (https://osf.io/hvy87/?view_only=95fa206d9e184f62b0ab0987fe907019), however, the hypothesis that the CA relationship would be stronger if only participants who indicated they would be able to make an identification were included in the analyses could not be evaluated. In Experiments 1 and 2, 96% and 97% of participants, respectively, indicated they would be able to make an identification.

participants was primarily female (.58; .001 indicated other and .004 did not respond) with a mean age of 38.64 years (*SD* = 12.89, *Range* = 18-82; .03 of participants chose not to respond or did not provide information about their age).

Design. The experiment used a between-subjects design and involved a manipulation of Target presence such that participants viewed a lineup containing the actor from the mock crime (target-present) or not (target-absent) and a manipulation of Target such that each mock-crime video featured one of four actors. The manipulation of interest was confidence query—the way that participants were first asked to report their confidence: using only a scale [*scale only*], following the approach recommended in the U.S.A of asking for confidence via a free report verbal statement in the eyewitness' own words [*own words*], following the approach in Scotland wherein participants are asked why they made the decision they did [*why*], or by selecting from a series of statements previously obtained from real witnesses [*selection*] (Behrman & Richards, 2005)[2]. These manipulations are described in further detail below.

Because very little research has used CAC curves and none has used inferential testing to compare them, sample size was determined based on the planned calibration analyses. Samples per cell for these analyses have varied from about 35 to over 350. Where the focus of the experiment is calibration, researchers in recent years have used 100 to 150 per cell, although Juslin, et al. (1996) suggested 200 per cell as best practice. The key manipulation, confidence query, had four levels, therefore a sample of 800 participants would be advisable. To account for loss of participants as a function of failing attention checks, 1000 participants were sought for Experiment 1 (final sample *N* = 912).

Materials. The experiment was conducted online using Qualtrics.

[2] This condition was included in order to determine the extent to which the CA relationship is maintained if eyewitnesses are asked to judge their confidence using a confidence lexicon (e.g., Ho, Budescu, Dhami, & Mandel, 2015). This condition is not central to the argument in this paper so it is referred to only briefly.

Mock crime. The four mock-crime videos used were selected from the set from which Mansour, Lindsay, Brewer, & Munhall (2009) selected their videos (not all were used by them but were part of the larger set available for use). All of the mock-crimes depict the same event with a different actor. The actor enters an office, speaks with someone off-screen, and when that person apparently leaves the room, rummages through a purse and steals money from it. The off-screen person then apparently returns; the actor speaks with her and then exits. Each video is approximately 30 seconds long with the actor visible throughout. The actors were all male, approximately 20 to 25 years old, and had brown hair.

Intervening task. Participants were presented with a portion of a Where's Waldo[3] image and a question (e.g., how many people are holding hands?). They were given 10 seconds to respond, otherwise they were presented another question. A total of 24 questions were used and the task continued until all questions had been presented. As many of the questions went unanswered, it was clearly a challenging task. Participants spent a mean of 3.67 minutes (*SD* = 0.59, *Range* = 0.13-4.42) on the task.

Lineups. The six-person lineups used were also selected from the set used by Mansour et al. (2009; see their Figure 1). They were presented with fair lineup instructions (Malpass & Devine, 1981). The lineups were constructed using the match-to-description approach (Luus & Wells, 1991), presented simultaneously, and participants were provided the option to select "not there" or one of the lineup members. Just the head and neck of lineup members were shown on a solid colour background; none of the lineup members had distinguishing marks, or accessories such as glasses or earrings.

Following Oriet & Fitzgerald (2018), targets similar in appearance to each other were selected so that they could act as innocent suspects to each other and thereby provide a means by which to have target-absent lineups constructed for the suspect. Thus, each target was

[3] TM & © 2008 Entertainment Rights Distribution Limited. All rights reserved.

paired with another target. For example, target 114 was paired with target 173 such that participants who saw the mock-crime video featuring target 114 later saw either the previously constructed target-present lineup for target 114 or the previously constructed target-present lineup for target 173 *as the target-absent lineup for target 114.*

Confidence. After selecting a lineup member or rejecting the lineup, participants were immediately asked about their confidence in their decision.

In the *scale only* confidence query condition, participants saw only the following question about their confidence: "Please tell us how confident you are in the accuracy of your lineup decision on this scale from 0% (not at all confident) to 100% (completely confident)." Below the question, participants were presented with a slider that could be moved anywhere along a line anchored at 0 and 100. The integer value between 0 and 100 associated with the current location of the slider was displayed when the participant held the slider by holding down the left mouse button.

In the *own words* confidence query condition, participants saw the following: "Please tell us how confident you are in the accuracy of your lineup decision in your own words." Participants could enter numbers or text in a box immediately below. When participants in this condition clicked the next button, they then received the same confidence question as those in the scale only condition.

In the *why* confidence query condition, participants saw the following: "If you chose someone from the lineup, please tell us why you have picked this person. If you responded, 'not there' to the lineup, please tell us why you did not pick anyone." As in the own words condition, participants were provided a text area immediately below and after responding, indicated their confidence on the 0-100% scale.

Finally, in the *selection* confidence query condition, participants saw the following: "Please click on the response that best represents your confidence in your lineup decision. If

you responded, "not there" to the lineup, please note that some of the options will not be relevant." Below this text, participants were presented with 34 statements in a 3 x 12 matrix. The statements corresponded with those collected by Behrman and Richards (2005) from real participants. The text was altered slightly to ensure it was appropriate for participants. Supplemental Table 1 lists these statements for comparison with Behrman and Richards.

Attention checks. Participants were asked two multiple choice questions to confirm they had paid attention to the mock-crime video: "What reason did the criminal give for coming to the office? [To pick up a VCR]" and "What did the criminal steal from? [A handbag]". The data from participants who incorrectly responded to both questions were excluded from analysis. Participants were also asked four questions to determine how much attention they had paid to the lineup instructions. Specifically, they were asked whether they had been told the criminal's appearance may have changed since the video, that the criminal may or may not be present in the lineup, that it is just as important to clear innocent persons from suspicion as to identify the guilty, and that they would have a certain amount of time to make a decision. The correct answer was yes only to the second and third questions. All participants were included regardless of their responses to these four questions, however, as it is plausible that real eyewitnesses may also sometimes fail to appreciate or remember some of the instructions they receive.

Procedure. After consenting to participate, participants viewed the mock-crime video. It was then explained that they were now eyewitnesses and they were asked whether they would be able to identify the mock-crime video actor from a lineup. They then completed the intervening task. Next, participants viewed fair lineup instructions, responded to the lineup, provided the confidence judgment associated with their condition (scale only, own words, why, or selection) and then, if they were in any condition other than the scale only condition, provided a confidence judgment on the 0-100% scale. The participants then

answered the attention check questions and were asked to report their age and sex. Finally, participants were debriefed, thanked, and provided a code that allowed them to be credited for their participation.

Coding and measures.

Identification decisions. Lineup decisions were coded as correct (suspect identifications from target-present lineups, rejections of target-absent lineups) or incorrect (filler identifications, suspect identifications from target-absent lineups). Inferential analyses on identification decisions were conducted with and without target-present filler identifications; where the nature of the results differ, the result excluding target-present filler identifications is reported in a footnote, otherwise all results include target-present filler identifications. Additionally, a choosing variable was created such that each lineup decision was coded for whether it was an identification (suspect identification or filler identification) or rejection.

Numeric confidence judgements. Sometimes when asked for confidence in their own words or why they made their decision, participants provided numeric information. Numeric statements in these cases and scale ratings of less than 50% were coded as low confidence (in line with the conclusions of Brewer, Keast, & Rishworth, 2002 and Brewer & Wells, 2006, that confidence ratings of below 50% are likely not predictive of accuracy). Numeric values of 50-79% were coded as medium confidence and 80% and above as high confidence[4]. These categories also align with Behrman and Richards (2005), whose approach was followed for coding the verbal confidence judgments.

Verbal confidence judgments. A review of the statements obtained in the why condition indicated that the reasons provided for lineup decisions were commonly judgments

[4] Prior literature has commonly categorized identifiers who are 90-100% confident as highly confident (e.g., Carlson, et al., 2016; Jalava et al., under review; Mickes, 2015; Semmler, et al., 2018; Wixted, et al., 2016; Wixted et al., 2015)

of confidence or references to features of the person remembered and/or identified. Thus, statements in the why condition were coded using the same scheme as the own words and selection conditions and were also coded for the presence/absence of information about features. Verbal judgements of confidence were coded as low, medium, or high confidence according to Behrman and Richards'(2005) coding scheme (see their Figure 1). Where participants made statements synonymous with statements indicated by Behrman and Richards, these were counted in the relevant category. For example, "completely confident" was coded as an instance of "absolutely certain." More ambiguous statements such as "I'm somewhat confident" were not coded as fitting into a most similar category as it was not clear what that category should be (e.g., "I am not exactly certain" [low confidence] or "Moderately sure" [medium confidence]).

Indeed, many of the obtained statements, particularly in the why condition, did not correspond clearly with Behrman and Richards' (2005) coding scheme. These statements were compiled and an independent set of raters (N = 36) rated each confidence judgment on the 0-10 scale also used by Behrman (2004, as cited in Behrman & Richards) to develop his coding scheme. In most cases the specific statements provided by participants were presented; however, where a number of participants had made nearly identical statements (e.g., a number of people stated the person they chose *looked just like* the man in the video because…. [remaining content varied]), only a single statement was presented as these statements were judged to indicate similar levels of confidence. The mean confidence rating given to the statements determined its categorization (see Supplemental Table 2).

Behrman and Richards categorised statements with a rating of 0-4 as low confidence 5-7 as medium confidence, and 8-10 as high confidence, based on the evaluations of a set of raters for what numeric values indicated low, medium, and high confidence. Given that I obtained means between categories, I rounded means to their nearest integer to determine

category placement (e.g., a mean rating of 4.56 was rounded to 5). Verbal statements in relation to rejections were coded by assigning them to the category which used the most similar language. For example, a statement like "No one looked like the perpetrator" was coded as medium confidence because the statement "looked like" was coded as medium according to Behrman and Richards' coding scheme.

Two raters coded all of the verbal statements obtained in the own words and why conditions. Percent agreement was high for the own words condition (93%) and for whether a feature was referred to in the why condition (92%), however, percent agreement was lower for confidence level in the why condition statements (74%), likely owing to the depth and variability in the statements made. These cases were commonly lineup rejections and sometimes reflected the fact that at times participants made statements consistent with multiple categories. For all three coded variables, I settled disagreements via discussion with the raters.

CACs and calibration. I constructed CAC and calibration curves as well as calibration statistics (calibration [C], over/underconfidence [O/U], discrimination [ANDI]) for each type of confidence query. Full details of how these were calculated can be found in the supplemental materials. The designated innocent suspects were identified less frequently than expected by chance (see Table 1); therefore, the false identification rate was estimated by dividing the target-absent lineup identifications by the nominal lineup size (i.e. six). Note that calibration curves are differentiated from the CAC curves by virtue of the fact that they use the target-absent identification rate without dividing it by nominal lineup size (i.e., all target-absent identifications are treated as errors, not just innocent suspect identifications); this allows comparison of calibration levels for identifications and rejections. As is common procedure (e.g., Brewer & Wells, 2006), filler identifications from target-present lineups were not used in constructing the calibration curves.

Results

Attention checks. In the usable sample (n = 968), .06 of participants (n = 56) responded incorrectly to both questions about the mock crime video and were therefore excluded from further data analyses. Thus, the analysed subset of data comprised 912 participants. For the four questions about the lineup instructions, the proportion of correct responses was .84 to the question about appearance change, .79 to the "may or may appear" question, .64 to the question about clearing the innocent/convicting the guilty, and .82 to the question about the time restriction.

Identification accuracy. Accuracy varied across targets, however, I collapsed across the targets for the analyses because it was hoped that the heterogeneity of responding would provide variability in confidence ratings. Table 1 provides details of accuracy as a function of target. Overall, participants who viewed target-present lineups made .50 correct identifications (*Range* = .31 - .71), .24 filler identifications (*Range* = .10 - .42), and .26 incorrect rejections (*Range* = .18 - .30). Participants who viewed target-absent lineups made .06 innocent suspect identifications (*Range* = .03 - .09), .26 filler identifications (*Range* = .14 - .36), and .68 correct rejections (*Range* = .60 - .76).

Preferred confidence query.

CAC curves. Figure 1 depicts the CAC curves for participants' first confidence judgment. Comparison of these curves suggests the own words condition is preferable as it is the only one for which low confidence judgments were associated with low accuracy. The why and selection conditions resulted in a much greater proportion of medium confidence compared to low or high confidence identifications. This preponderance of medium confidence judgments likely reflects that it was challenging to interpret statements in the why condition and that participants avoided extreme statements in the selection condition.

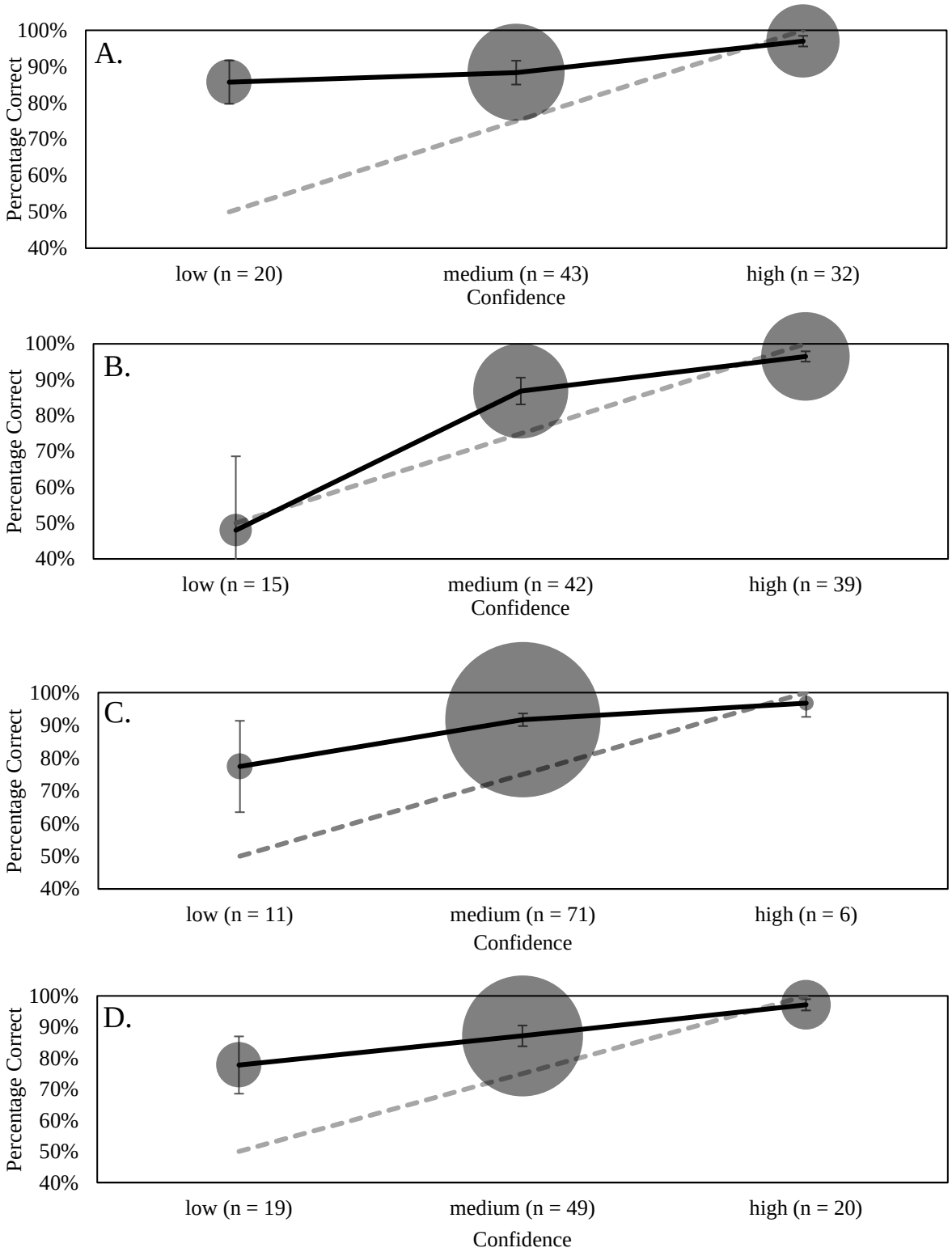


Figure 1. Confidence-Accuracy Characteristic (CAC) curves for identifications in the A) scale only, B) own words, C) why, and D) selection conditions in Experiment 1. The size of the plotted data point reflects the proportion of cases that contributed to the calculation (as suggested by Evelo, Lee, Modjadidi, & Penrod, 2018). Standard error bars were estimated using 10,000 bootstrapped samples.

Somewhat surprisingly, the scale only approach did not show a particularly strong relationship across all confidence categories. Figure 2 provides CACs for the scale ratings for all conditions using three categories and using five categories (B) in the scale only condition and (C) the full sample. There is considerable variability in the CA relationship when confidence is below 60%, as has been found elsewhere (e.g., Brewer & Wells, 2006).

Logistic regression. Logistic regression was used to examine whether one type of confidence query was preferable to the others. Type of confidence query (scale only, own words, why, or selection), confidence level (low, medium, and high), and their interaction were entered as predictors of accuracy using backwards stepwise (likelihood ratio) entry. Contrasts were set to compare each condition to the scale only condition since this is arguably the control condition in this experiment (i.e., it is the approach commonly used in research and for which we have the most data). For identifications, the final model was significant but only confidence level was maintained in the model, $\chi^2(1, n = 475) = 41.34$, $p < .001$, $r_{\text{Nagelkerke}} = .11$; participants were 2.55 times more likely to be accurate for each increase in confidence. For rejections, the final model also contained only confidence level, $\chi^2(1, n = 434) = 13.37$, $p < .001$, $r_{\text{Nagelkerke}} = .04$, such that participants were 1.80 times more likely to be accurate for each increase in confidence level. Thus, none of the approaches to collecting confidence was significantly better or worse than the scale only approach.

Predictive ability of different confidence queries. Logistic regression was used to separately assess the extent to which confidence level (low, medium, high) predicted the accuracy of identifications and rejections for each type of confidence query. For each, a separate logistic regression was conducted for 1) identification accuracy and 2) rejection accuracy with confidence level (low, medium, high) as the predictor.

For the scale only condition, confidence level was a significant predictor of identification accuracy, $\chi^2(1, n = 123) = 6.91$, $p = .009$, $r_{\text{Nagelkerke}} = .07$. For each increment in

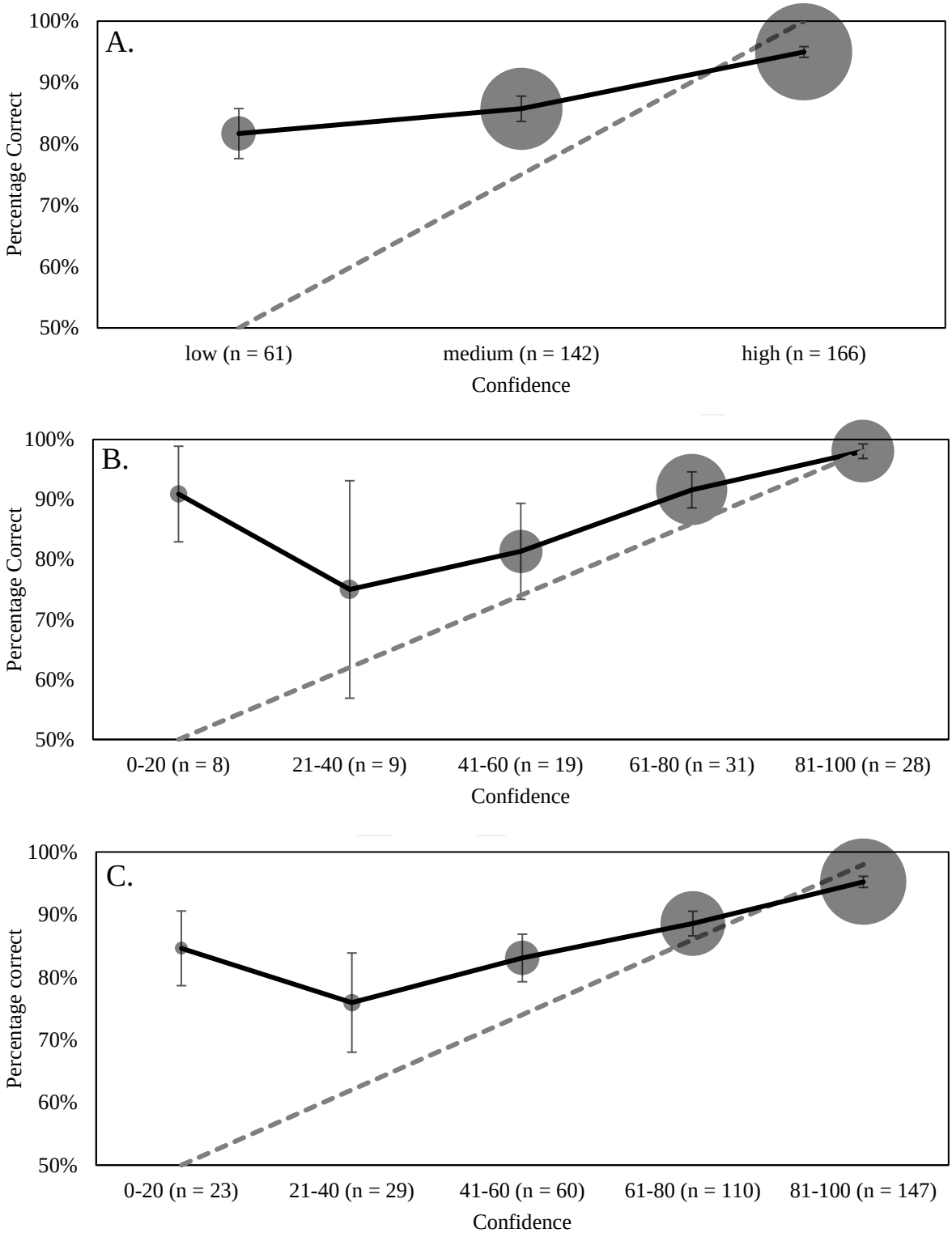


Figure 2: Confidence-Accuracy Characteristic (CAC) curves for scale ratings A) across all participants using three bins, B) the scale only condition using five bins, and C) all participants using five bins in Experiment 1. The size of the plotted data point reflects the proportion of cases that contributed to the calculation (as suggested by Evelo, et al., 2018). Standard error bars were estimated using 10,000 bootstrapped samples.

confidence level, participants were 1.95 times more likely to make a correct than incorrect identification. Correct categorizations improved from 50.4% to 61.8% when confidence was included in the model. Scale ratings of confidence also predicted the accuracy of rejections, $\chi^2(1, n = 94) = 7.13$, $p = .008$, $r_{\text{Nagelkerke}} = .12$. For each increment in confidence level, participants were 2.56 times more likely to correctly versus incorrectly reject the lineup, however correct categorizations were maintained at 79.8%.

For the own words condition, participants were 3.90 times more likely to be accurate for each increment in their confidence level, $\chi^2(1, n = 125) = 24.14$, $p < .001$, $r_{\text{Nagelkerke}} = .24$, for identifications. The model improved correct categorization of identifications from 55.2% to 69.6%. There was no relationship between confidence level and rejections, $\chi^2(1, n = 110) = 0.84$, $p = .36$, $r_{\text{Nagelkerke}} = .01$, such that assuming all decisions were correct led to 71.8% accurate categorizations and did not improve when confidence was incorporated as a predictor.

For the selection condition, confidence level was a significant predictor of identification accuracy, $\chi^2(1, n = 112) = 9.53$, $p = .002$, $r_{\text{Nagelkerke}} = .11$. Participants were 2.47 times more likely to make a correct identification for each increment in their confidence level. Categorization accuracy improved from 55.4% to 63.4% when confidence was used as a predictor. Confidence was not a significant predictor of the accuracy of rejections in this condition, $\chi^2(1, n = 114) = 3.72$, $p = .054$, $r_{\text{Nagelkerke}} = .05$, $OR = 1.82$, with categorization maintained at 69.3% with the use of confidence as a predictor.

For the why condition, two predictors were entered—confidence level and whether the participant referred to features or not in their response. For identifications, the model was significant, $\chi^2(2, n = 115) = 6.29$, $p = .04$, $r_{\text{Nagelkerke}} = .07$. Only confidence level was a significant predictor of the accuracy of identifications, $\chi^2(1) = 4.86$, $p = .025$, $OR = 2.98$; the presence/absence of information about features was not a significant predictor ($p = .32$).

Classification of identification decisions improved from 52.2% to 59.1% using the model with both predictors. The model for rejections was not significant, $\chi^2(2, n = 116) = 2.94$, $p = .23$, $r_{Nagelkerke} = .04$ and neither predictor approached significance (all ps > .11); classification accuracy improved from 72.4% to 73.3% when the predictors were included.

In summary, the results were consistent with the hypothesis that scale ratings and certain verbal judgments of confidence would predict identification accuracy. However, contrary to expectations, asking why the participant made their decision predicted identification accuracy similarly to the other confidence query conditions.

Calibration. Figure 3A depicts the calibration curves for identifications and rejections. Calibration was similar for identifications and rejections when confidence was above 60% and 20% or lower but better for identifications than rejections when confidence was between 21% and 60%. However, rejections evidenced underconfidence whereas identifications evidenced the typical pattern of overconfidence. These results are consistent with the expectation of a higher CA relationship for identifications than rejections and that highly confident rejections are associated with accuracy. Interestingly, calibration for identifications was strong for ratings of 20% and higher.

Calibration statistics for identifications (see Table 2) were evaluated with inferential confidence intervals (ICIs, see Table 3) to compare scale ratings by those who provided a scale rating only to those that provided a scale rating after a verbal response[5]. All possible comparisons were conducted with a Bonferroni correction. Only one difference was significant: the own words condition resulted in better calibration than the why condition.

[5] The data and R-code used to compute ICIs and relevant parameters for both experiments can be found on the OSF project page: https://osf.io/qydb4/?view_only=c9038f945ca44023bad61ad37af5826f

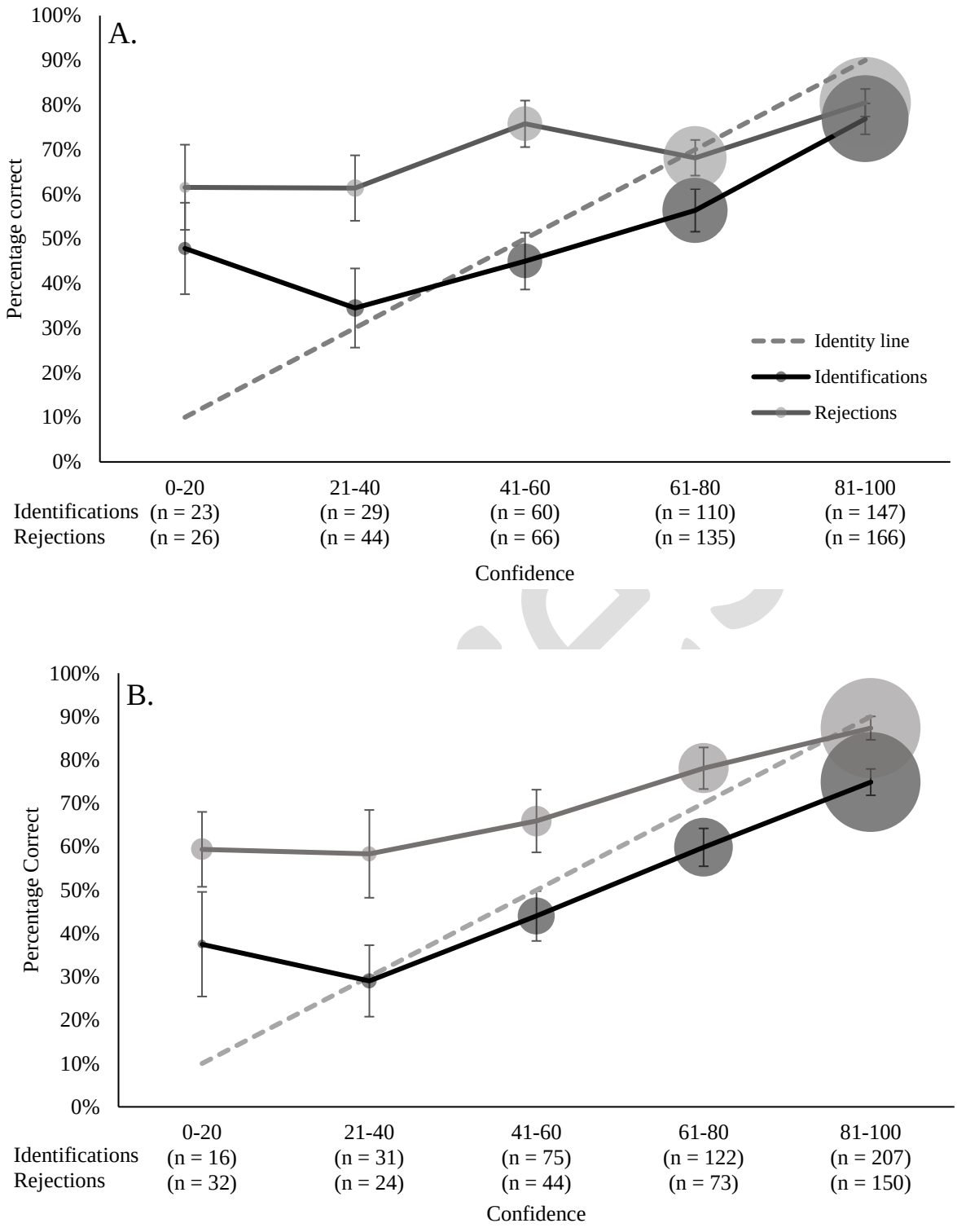


Figure 3: Calibration curves for scale ratings made by all participants in A) Experiment 1 and B) Experiment 2. The size of the plotted data point reflects the proportion of cases that contributed to the calculation (as suggested by Evelo, et al., 2018). Standard error bars were estimated using 10,000 bootstrapped samples.

ICIs were also used to compare the calibration statistics when only a scale judgement was obtained to a scale judgement obtained after a verbal judgment (i.e., collapsing across the own words, why, and selection conditions; see Table 3). The scale only condition resulted in less overconfidence than the other conditions but calibration and discrimination did not differ. An analysis of variance (ANOVA; see supplemental materials) indicated that following the why or selection conditions with a scale rating reduced confidence compared to only a scale judgment.

Correspondence between verbal judgments and scale ratings of confidence. Table 4 illustrates the correspondence between participants' initial confidence judgments and their subsequent scale rating when they made an identification. Statements coded as indicating high confidence were generally associated with scale ratings indicative of high confidence (> .90 of the statements); however, the correspondence was much lower for statements coded as indicating medium (.44-.75) or low confidence (.25-.79).

Correlations. As expected, the CA correlation was stronger for identifications, $r(473) = .29$, $p < .001$, than rejections, $r(432) = .17$, $p < .001$, using the first-collected confidence measure (low, medium, high). Using the scale ratings provided by all participants, the CA correlation was also higher for identifications, $r(475) = .27$, $p < .001$ than for rejections, $r(433) = .14$, $p = .005$.

Discussion

The results were broadly as hypothesised. The own words and selection conditions performed similarly to the scale only condition for all measures. Unexpectedly, the why condition resulted in a similar CA relationship to the other conditions. One reason may be that many participants provided a confidence judgement as well as or instead of an explanation for their decision (see Supplemental Table 3). However, scale ratings in the why condition resulted in greater overconfidence than in the own words condition and the

approach resulted in a preponderance of medium confidence judgments. Finally, and as expected, the CA relationship was stronger for identifications than rejections but high-confidence rejections were highly likely to be accurate.

Across the verbal confidence conditions, the coding of verbal statements corresponded well with scale-based confidence only for high-confidence judgments, suggesting caution when interpreting verbal statements. Indeed, there was considerable variability in the numeric ratings assigned these verbal statements by independent participants.

Experiment 1 demonstrated that verbal statements and numeric scale approaches to confidence predict identification accuracy for simultaneous lineups. However, many police departments utilize sequential lineups. Experiment 2 examined the CA relationship for own words and scale only confidence judgements using sequential lineups.

Experiment 2

In Experiment 2 two confidence queries were tested, participants viewed sequential lineups, and different target-innocent suspect pairs were used. The hypotheses were the same as in Experiment 1 (where relevant).[6]

Method

Participants. Participants were again recruited through Amazon Mechanical Turk/Turk Prime, using the same requirements as in Experiment 1. Participants who did not complete the experiment (n = 242) and completed the experiment multiple times (based on IP addresses; n = 39/49 cases) were dropped from the sample. The usable sample of participants (N = 981) was approximately half female (.54; .003 indicated other and .02 did not respond) with a mean age of 37.73 years (SD = 12.77, $Range$ = 18-80; .07 of participants chose not to respond or did not provide information about their age).

[6] The preregistration can be found here: https://osf.io/2taec/?view_only=2ff6d3ba9ac74de6be06749645c0073d

Design. As in Experiment 1, Target presence, Target, and Confidence query were manipulated. Only two levels of confidence query were used, though: scale only or own words (the own words condition arguably performed the best of the verbal approaches tested in Experiment 1 and is widely used). Experiment 1 revealed quite small subsamples in the low confidence category, therefore double the sample size per cell was sought for Experiment 2. Experiment 2 had 16 conditions (2 Target presence x 4 Target x 2 Confidence Query) but the key manipulation of confidence query had only two levels, therefore 1000 participants were sought.

Materials. The materials used in Experiment 2 were the same as Experiment 1 except that effort was made to select more confusable targets in order to raise the innocent suspect identification rate somewhat. All targets in the Mansour et al. (2009) set matched the same general description and in Experiment 1 and therefore were broadly similar looking; no attempt was made to choose pairs of targets that were similar to each other. However, for Experiment 2, I inspected the targets in the stimulus set and chose the two pairs of targets I felt were the most similar to each other in the set.

In addition, the lineups were presented sequentially following the, perhaps, typical procedure used in the U.S.A. That is, participants viewed all lineup members one after another, could select more than one lineup member (although they were not informed of this fact in the lineup instructions), and could view the lineup more than once if they chose (i.e., after viewing the last lineup member, participants were asked if they would like to view the lineup again and could do so up to five times).

After responding "no" to the question about viewing the lineup again or after viewing the lineup five times, if participants had selected more than one lineup member, they were asked to choose only one lineup member and shown the lineup members sequentially once more. Participants who did not choose a single lineup member in this step were excluded

from analyses for practical reasons. For the analyses, the final selection was examined as this decision would be the one expected to be considered by practitioners. Any time a participant responded "yes" to a lineup member, they were asked to indicate their confidence before viewing another lineup member; if a participant responded "no" to all lineup members, they were asked to indicate their confidence at that point (prior to being asked if they wanted to view the lineup again). Suspects appeared in position three or five.

Procedure. The procedure was identical to Experiment 1 other than the differences due to using sequential rather than simultaneous lineups and different target/lineup pairs, described above.

Coding and measures. Identification decisions, numeric confidence judgments, CACs, calibration curves, and calibration statistics were coded as in Experiment 1. The verbal confidence judgments obtained in the own words condition were coded according to both Behrman and Richards' (2005) scheme and the one developed for Experiment 1. Percent agreement amongst the two coders (who coded all cases) was 92%; as in Experiment 1, I settled the disagreements amongst the coders.

Results

Attention checks. Of 981 usable participants, .12 (n = 105) were excluded because they responded incorrectly to both attention check questions, resulting in 876 participants. For the questions about the lineup instructions, the proportion of correct responses were .78 to the question about appearance change, .66 to the "may or may appear" question, .64 to the question about clearing the innocent/convicting the guilty, and .74 to the question about the time restriction.

Identification accuracy. I again collapsed across targets but provide the identification decision rates by target in Table 1. Approximately .11 of the sample selected more than one lineup member on their last view of the lineup before being asked to clarify

their choice (.09 selected two lineup members, .01 selected three; and .01 selected four, five, or all six lineup members). A small proportion (.02, *n* = 17) of participants selected multiple faces on the final, clarifying view of the lineup and therefore were not included in further analyses because it was unclear who they thought was the perpetrator. In the remaining sample (*n* = 859), .69 viewed the lineup once, .29 viewed it a second time, and .01 viewed it a third time. The distribution of multiple laps was nearly identical across the scale only (.30) and own words conditions (.31). Supplemental Table 4 illustrates the association between lineup laps and the selection of multiple lineup members.

I considered identification accuracy based on each eyewitness' final identification decision (i.e., on their final lap and after clarifying multiple identifications). The overall rate of correct identifications was .64 (*Range* = .59 - .70), of target-present filler identifications was .20 (*Range* = .12 - .28), and of incorrect rejections was .17 (*Range* = .13 - .24). For target-absent lineups, the rate of suspect identifications—higher than in Experiment 1 but not quite at chance, as hoped—was .12 (*Range* = .09 - .16), of filler identifications was .29 (*Range* = .16 - .42), and of correct rejections was .59 (*Range* = .50 - .70).

Preferred confidence query.

CAC curves. Figure 4 depicts the CAC curves for the scale only and own words conditions. Consistent with my expectations, the curves almost completely overlap for the medium and high confidence categories. Although the points for the low confidence categories appear to differ when participants who viewed the lineup only once are examined (Figure 4A), the large standard errors indicate no difference.

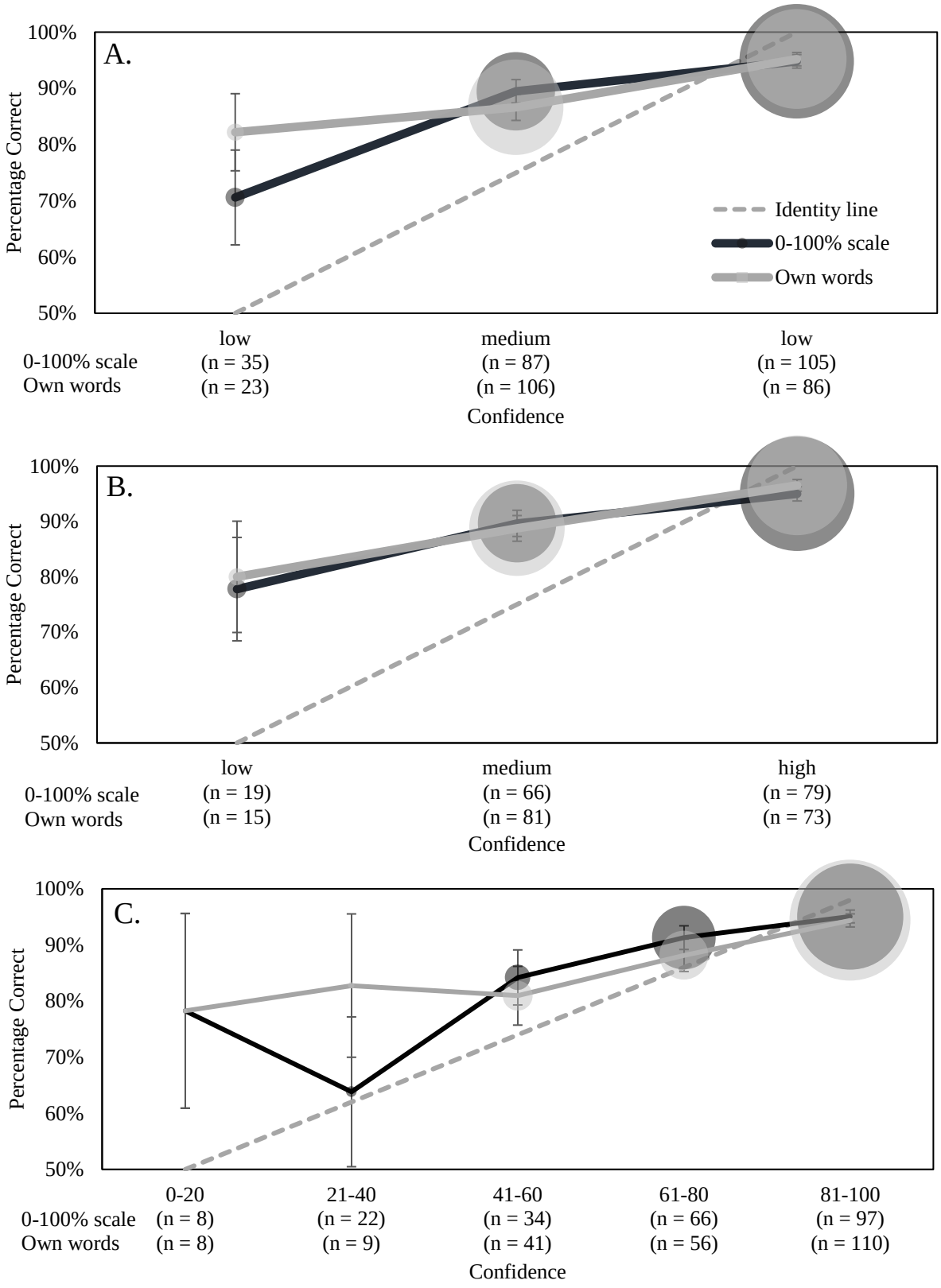


Figure 4. Experiment 2 Confidence-Accuracy Characteristic (CAC) curves for the initial confidence judgment (low, medium, or high) made by A) all participants or B) participants who viewed the lineup only once and C) the scale ratings for all participants. The size of the plotted data point reflects the proportion of cases that contributed to the calculation (as suggested by Evelo, et al., 2018). Standard error bars were estimated using 10,000 bootstrapped samples.

Logistic regression. Logistic regression was used to inferentially test whether one type of confidence query (scale only vs. own words) was more predictive of identification accuracy. Confidence query, confidence level (low, medium, high), and their interaction were entered as predictors. The model was significant, $\chi^2(3, n = 527) = 38.91$, $p < .001$, $r_{\text{Nagelkerke}} = .10$, however, neither confidence query ($p = .91$), nor the interaction of confidence query and confidence level ($p = .86$) were significant. Only confidence level was significant, $\chi^2(1) = 22.03$, $p < .001$, $OR = 2.36$, such that for each one step increase in confidence level, participants were 2.36 times more likely to be accurate. The predictor model improved classification accuracy from 51.4% to 62.2%. The model was repeated using only suspect identifications and the conclusions were the same, however, confidence was a stronger predictor: for each increase in confidence, the identification was 3.20 times more likely to be correct. Thus, as hypothesised, confidence query did not interact with confidence level.

I entered the same predictors into a logistic regression model predicting rejection accuracy and again found a significant model, $\chi^2(3, n = 321) = 21.02$, $p < .001$, $r_{\text{Nagelkerke}} = .10$. As with the model for identifications, only confidence level was a significant predictor, $\chi^2(1) = 10.43$, $p = .001$, $OR = 2.33$; however, the rate of correct classifications was static at 77.3%. To conclude, as expected, confidence level predicted identification accuracy regardless of confidence query, but also predicted rejection accuracy.

Predictive ability of different confidence queries. To determine the effectiveness of each method of requesting confidence, I conducted separate logistic regressions with confidence level as the predictor of 1) identifications and 2) rejections for the scale only and own words conditions.

For identifications in the scale only condition, confidence level significantly predicted accuracy, $\chi^2(1, n = 269) = 24.08$, $p < .001$, $r_{\text{Nagelkerke}} = .11$. Each increase in confidence meant a participant was 2.36 times more likely to be accurate. Inclusion of confidence level as a

predictor improved classification accuracy from 52.0% to 61.7%. For rejections in the scale only condition, confidence level was also a significant predictor of accuracy, $\chi^2(1, n = 161) = 10.75$, $p = .001$, $r_{\text{Nagelkerke}} = .10$. For each one step increase in confidence, participants were 2.33 times more likely to be accurate. Classification accuracy was 81.4%, regardless of whether confidence level was included in the model, however.

When participants in the own words condition identified someone from the lineup, confidence level significantly predicted their accuracy, $\chi^2(1, n = 258) = 15.75$, $p < .001$, $r_{\text{Nagelkerke}} = .08$[7]. Each increase in confidence level meant a participant was 2.24 times more likely to be accurate. Classification accuracy improved from 50.8% to 62.8% with the inclusion of confidence level as a predictor. For rejections in the own words condition, confidence level was also a significant predictor of accuracy, $\chi^2(1, n = 160) = 7.16$, $p = .01$, $r_{\text{Nagelkerke}} = .06$, such that for each increase in confidence level, participants were 1.93 times more likely to be accurate. Classification accuracy did not improve from 73.1%, however.

Calibration. Figure 3B depicts the calibration curves for identifications and rejections in Experiment 2. The figure indicates that, as expected, the CA relationship is stronger for identifications than rejections, although high confidence rejections were associated with accuracy. Similar to Experiment 1, identifications and rejections were relatively well-calibrated for the upper two confidence bins (confidence greater than 60%) and poorly calibrated for the lowest confidence bin (20% or less). Also, like Experiment 1, calibration for identifications was very good for selections but poorer for rejections for the middle two confidence bins (21% to 60% confidence).

Again, calibration statistics (see Table 2) were evaluated using inferential confidence intervals (see Table 3) in order to compare the scale ratings given by all participants as a

[7] If only suspect identifications are included in the analysis, confidence is no longer a significant predictor ($p = .14$), likely owing to the small number of cases ($n = 151$) and the low proportion of innocent suspect identifications (20 cases; .12).

function of whether they gave only the scale rating or first stated their confidence in their own words. Consistent with my hypotheses and Experiment 1, providing a scale rating after a verbal judgment did not affect the utility of the scale rating. However, an ANOVA indicated that providing only a scale judgement led to lower confidence ratings than when the scale judgement followed an own words judgment (see supplemental materials), contrary to Experiment 1 where no difference was found.

Correspondence between verbal judgments and scale ratings of confidence. Table 4 illustrates the correspondence between verbal judgments (in the own words condition) and participants' subsequent scale ratings when they made an identification. As in Experiment 1, correspondence between verbal judgments and scale ratings was strong in the high-confidence category (.95) but weaker in the low- (.69) and medium-confidence categories (.63).

Correlations. The CA correlation for the first-collected confidence measure (low, medium, high) for identifications, $r(440) = .29$, $p < .001$, was similar to rejections, $r(319) = .25$, $p < .001$. This relationship held when only scale ratings for identifications, $r(449) = .30$, $p < .001$, and rejections, $r(321) = .25$, $p < .001$, were examined. These results are contrary to my expectations—participants in the current experiment had an unexpectedly strong CA relationship for rejections.

Discussion

Experiment 2 extends key findings from Experiment 1 to sequential lineups. Own word judgments and scale ratings similarly predicted accuracy, as hypothesised. Replicating Experiment 1, the correspondence between a verbal confidence judgment and a scale rating was strong only for high-confidence judgments and high-confidence rejections were associated with high accuracy. However, in this experiment, the CA relationship for rejections was nearly equivalent to that of identifications. This unexpected result may reflect

the experimental procedure. Participants were offered the option to view the lineup again (up to five times). When an eyewitness views an entire lineup without seeing the culprit, the feeling that they may have missed the culprit may reduce their confidence and thus the CA relationship. A second view of the lineup may reaffirm a rejection decision thus increasing the CA relationship. Indeed, of the 265 participants that viewed the lineup more than once, nearly half (113) initially selected no one and maintained that decision (see Supplemental Table 4). Unfortunately, the benefit would come at too high a cost: multiple laps of target-absent sequential lineups increases incorrect identifications (Horry, Brewer, Palmer, & Weber, 2015).

General Discussion

These data replicate prior findings that eyewitness confidence can be predictive of accuracy and extend this conclusion to approaches used in practice. Identification confidence in the participant-eyewitness' own words (U.S.A.) or when asked why they made their identification (Scotland) was as predictive of accuracy as on a numeric scale (categorized as low, medium, or high). A similar CA relationship was found for suspect versus filler identifications as in a sample of real eyewitnesses (see supplemental analyses), suggesting these findings may generalize to real eyewitnesses. However, the results point to important considerations for the use of confidence as a predictor of identification accuracy (see also Sauer et al., 2019).

First, the conclusions differed for simultaneous and sequential lineups. For simultaneous lineups, the CAC curves suggest that an own words judgment better predicts accuracy than a scale judgement because of the strong CA relationship for low confidence judgements in the former condition; for sequential lineups, the two approaches elicited a similar CA relationship. As Experiment 1 and 2 used different stimuli and the sample sizes were quite different, inferentially comparing the simultaneous and sequential lineups is

inappropriate; however, the data suggest confidence query may interact with identification procedure.

Second, these data highlight the considerable challenges in using verbal judgements of confidence. Requesting confidence in the participant-eyewitness' own words or asking why they made their identification produced a good CA relationship but coding the verbal statements was challenging—they were highly variable (see Supplementary Table 2) as has been found in other fields (e.g., Dhami & Wallsten, 2005). Agreement was lower among coders in the why than the own words condition (74% vs. 93%) and confidence judgments concentrated more heavily in the medium category in the why condition (Figure 1C), indicating the own words approach is preferable—but both approaches will be challenging to validate for practice.

As Table 4 demonstrates and prior research has shown (e.g., Brun & Teigen, 1988; Budescu, et al., 1988; Dhami & Wallsten, 2005), verbal confidence judgments are often interpreted differently from intended. Independent participants rated the collected statements (0-10) and the range for many statements spanned nearly the entire scale (i.e., the large standard deviations in Supplemental Table 2), suggesting broad membership functions for specific phrases, as has been found in other fields (e.g., Dhami & Wallsten, 2005; Ho, et al., 2015). Karelitz and Budescu (2004) developed a method for comparing the meaning of probability phrases across individuals but the approach is time consuming and therefore not feasible for the criminal justice system.

Meaningful boundaries for what constitutes low, medium, and high confidence are needed. In the absence of such boundaries, police officers and prosecutors must interpret for themselves. A police officer may decide whether to continue to pursue a suspect or not based on the eyewitness' confidence, a prosecutor likely considers the eyewitness' confidence in deciding whether to prosecute a case, and judges and jurors interpret an eyewitness'

confidence in order to determine how much weight to give to the eyewitness' evidence. In addition to the wide variability in possible interpretations applied to verbal phrases, beliefs about confidence and/or the weight of other evidence at each point are likely to influence how identification evidence is treated (Brun & Teigen, 1988; Hasel & Kassin, 2009; Wallsten, Fillenbaum, & Cox, 1986; Weber & Hilton, 1990). Moreover, people underestimate the variability in the interpretation of verbal probability statements. Brun and Teigen (1988) asked participants to estimate the range of numeric probabilities covered by a phrase for 90% of the population. The mean range varied from .5-.75 of the actual range, empirically determined in the sample. Further muddying the issue is the finding that different words carry a different emotional charge, which may affect how they are interpreted (Brun & Teigen, 1988).

Given that extant research demonstrates a strong CA relationship for high confidence identifications and the relatively strong correspondence between verbal and scale-based judgements when confidence was high, one may suggest relying only on high confidence identifications as evidence of guilt. This approach would undoubtedly reduce wrongful convictions but also the likelihood of prosecution of a high number of cases where, if the eyewitness' confidence had been queried using an alternative approach (e.g., a numeric scale), the eyewitness may have appeared highly confident. Notably, there was a preponderance of medium confidence identifications when participants were asked why they made their identification. Furthermore, low and medium confidence identifications in the own words condition were surprisingly accurate (.86 and .89, respectively). Perhaps we should not dismiss eyewitnesses just because they are not highly confident. Likewise, we must be cautious with highly confident eyewitnesses (Sauer et al., 2019)—highly confident participants were highly but not perfectly accurate (.95-.98).

Not only do we require safeguards for the interpretation of verbal statements, but also

to ensure verbal statements are recorded faithfully. Variability in recording has implications for the apparent effectiveness of a procedure (Rodriguez & Berry, 2019; Steblay, 2011). Encouragingly, in the UK, officers are required to record all comments made by eyewitnesses in relation to their identification (PACE Code D, 2017, p. 39).

One way to reduce variability in intended meaning and interpretation is the development of a confidence lexicon. In fields where analysts provide judgments about the likelihood of events (e.g., climate change, intelligence), organizations have begun developing lexicons wherein verbal statements represent a range of numeric values (e.g., Ho, et al., 2015; Dhami, 2018). Asking eyewitnesses to express their confidence by selecting from a series of statements resulted in a similar CA relationship to other methods, providing preliminary support for such an approach; although this approach resulted in a preponderance of medium-confidence judgments. Nonetheless, this approach may be more palatable for those reluctant to request numeric confidence judgments and could reduce variability in interpretation. On the other hand, people are reluctant to use others' definitions of verbal phrases (Wallsten & Budescu, 1990) and as one prosecutor commented, this approach would be unacceptable to the courts because it entails "putting words in the eyewitness' mouth."

Perhaps eyewitnesses could provide a scale rating after a verbal statement. Obviously more nuanced information is obtained from a 0-100% scale than categorizing verbal statements as low, medium, or high confidence. The results suggest there may be little harm in this approach. No differences in scale ratings were found in calibration, over/underconfidence, or discrimination for the scale only compared to the verbal-first conditions. Scale ratings obtained solely versus after a verbal statement were lower suggesting a potential benefit to obtaining both judgements from the perspective of the prosecution—insofar as more confident eyewitnesses are seen as more reliable. However, the CA relationship did not differ across these two situations and in Experiment 1, a scale rating

after a verbal judgement resulted in greater overconfidence. Further study is needed.

A surprisingly strong CA relationship was found for rejections. The calibration curves indicate this result may be due to the high level of calibration for the upper-middle levels of confidence. Calibration for rejections is typically good only at the uppermost confidence levels (Wixted & Wells, 2017). However, here rejections were reasonably well calibrated for scale ratings above 60% (simultaneous lineups) and 40% (sequential lineups). Perhaps this finding reflects internal processes at work when participants provided a numeric rating after a verbal one or because participants were offered the opportunity to view sequential lineups again. Regardless, these findings highlight the need to better understand the circumstances under which confidence judgments are predictive of lineup decision accuracy.

There are limitations to this research. I aimed to make the sequential lineup procedure as analogous to real-life practice as possible but doing so introduced decision points that may not occur in practice. Participants were explicitly asked if they wanted to view the lineup again and could do so up to five times. Best practice guidance discourages this question; therefore, the CA relationship may appear stronger (or weaker) for sequential lineups here than it would otherwise. However, the CAC curves for participants who viewed the lineup once are similar to those for the entire sample (see Figure 4). A second limitation is in the translation of scale ratings to low, medium, or high confidence. Although I followed prior research (Behrman & Richards, 2005; Brewer & Wells, 2006; Wixted & Wells, 2017), the numeric boundaries used here and in the broader eyewitness identification literature have been somewhat arbitrarily established. An empirically-determined or otherwise systematic approach to these boundaries would improve validity and the comparability of findings across experiments. Third, these experiments used relatively uniform stimuli. Although the suspect identification rates were similar to practice (~50%; Behrman & Richards, 2005), real-world cases involve considerably more variability. Replications should examine how robust these

patterns are.

To summarise, these experiments link the way confidence is collected in experiments and practice. Across the methods tested—a 0-100% scale, asking for confidence in the eyewitness' own words, asking why the eyewitness made the decision they did, and asking the eyewitness to select the statement that best represented their confidence—confidence predicted accuracy. The challenges involved in interpreting verbal confidence statements are not small and only when confidence was high were verbal and scale judgments reliably interpreted similarly. More research is needed on the way confidence is collected verbally, including the development of benchmarks for low, medium, and high confidence and on how to validly obtain and then interpret the meanings of these statements.

Author Contributions

The first author conceived the research questions and hypotheses, designed the experiments, prepared and submitted the preregistration, partially programmed the experiments, coded all data except the verbal statements, analysed the data, and wrote and edited the manuscript.

The author would like to gratefully acknowledge the assistance of Taylor Ashby with the experiment programming, of Rhiannon Batstone with the preparation of the task completed by the independent participants and testing the experiment, and of Eilidh Haig, Kim Schneider, and Vasileios Sideropoulos with data coding and revision of the coding scheme.

References

Barnes, A. (2016). Making intelligence analysis more intelligent: Using numeric probabilities. *Intelligence and National Security, 31*, 327-344. doi: 10.1080/02684527.2014.994955

Brewer, N., Keast, A., & Rishworth, A. (2002). The confidence-accuracy relationship in eyewitness identification: The effects of reflection and disconfirmation on correlation and calibration. *Journal of Experimental Psychology: Applied, 8*, 44-56. doi: 10.1037/1076-898X.8.1.44

Brewer, N., & Wells, G. L. (2006). The confidence-accuracy relationship in eyewitness identification: effects of lineup instructions, foil similarity, and target-absent base rates. *Journal of Experimental Psychology: Applied, 12*, 11-30. doi: 10.1037/1076-898X.12.1.11

Brun, W., & Teigen, K. H. (1988). Verbal probabilities: Ambiguous, context-dependent, or both?. *Organizational Behavior and Human Decision Processes, 41*, 390-404. doi: 10.1016/0749-5978(88)90036-2

Budescu, D. V., & Wallsten, T. S. (1990). Dyadic decisions with numerical and verbal probabilities. *Organizational Behavior and Human Decision Processes, 46*, 240-263. doi: 10.1016/0749-5978(90)90031-4

Budescu, D. V., Weinberg, S., & Wallsten, T. S. (1988). Decisions based on numerically and verbally expressed uncertainties. *Journal of Experimental Psychology: Human Perception and Performance, 14*, 281-294. doi: 10.1037/0096-1523.14.2.281

Busey, T. A., Tunnicliff, J., Loftus, G. R., & Loftus, E. F. (2000). Accounts of the confidence-accuracy relation in recognition memory. *Psychonomic Bulletin & Review, 7*, 26-48. doi: 10.3758/BF03210724

Canty, A., & Ripley, B. (2019). boot: Bootstrap R (S-Plus) Functions. R package version 1.3-

22.

Carlson, C. A., Young, D. F., Weatherford, D. R., Carlson, M. A., Bednarz, J. E., & Jones, A. R. (2016). The influence of perpetrator exposure time and weapon presence/timing on eyewitness confidence and accuracy. *Applied Cognitive Psychology, 30*, 898-910. doi: 10.1002/acp.3275

Cutler, B. L., Penrod, S. D., & Dexter, H. R. (1990). Juror sensitivity to eyewitness identification evidence. *Law and Human Behavior, 14*, 185-191. doi: 10.1007/BF01062972

Cutler, B. L., Penrod, S. D., & Stuve, T. E. (1988). Juror decision making in eyewitness identification cases. *Law and Human Behavior, 12*, 41-55. doi: 10.1007/BF01064273

DiCiccio, T. J., & Efron, B. (1996). Bootstrap confidence intervals. *Statistical Science, 11*, 189-212.

Dhami, M. K. (2018). Towards an evidence-based approach to communicating uncertainty in intelligence analysis. *Intelligence and National Security, 33*, 257-272. doi:

Dhami, M. K., & Wallsten, T. S. (2005). Interpersonal comparison of subjective probabilities: Toward translating linguistic probabilities. *Memory & Cognition, 33*, 1057-1068. doi: 10.3758/BF03193213

Dodson, C. S., & Dobolyi, D. G. (2016). Confidence and Eyewitness Identifications: The Cross- Race Effect, Decision Time and Accuracy. *Applied Cognitive Psychology, 30*, 113-125. doi: 10.1002/acp.3178

Ericsson, A. (2003). Valid and non-reactive verbalization of thoughts during performance of tasks towards a solution to the central problems of introspection as a source of scientific data. *Journal of Consciousness Studies, 10*, 1-18. doi:

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review, 87*, 215-251. doi: 10.1.1.697.3088

Ericsson, K. A., & Simon, H. A. (1998). How to study thinking in everyday life: Contrasting think-aloud protocols with descriptions and explanations of thinking. *Mind, Culture, and Activity, 5*, 178-186. doi: 10.1207/s15327884mca0503_3

Evelo, A., Lee, J., Modjadidi, K., & Penrod, S. (2018, July). *The role of lineup bias in witness accuracy, the confidence-accuracy relationship and the courtroom value of witness confidence.* Paper at the European Association for Psychology and Law. Turku, Finland.

Hall, L., Johansson, P., & Strandberg, T. (2012). Lifting the veil of morality: Choice blindness and attitude reversals on a self-transforming survey. *PloS one, 7*, e45457. doi: 10.1371/journal.pone.0045457

Hasel, L. E., & Kassin, S. M. (2009). On the presumption of evidentiary independence: Can confessions corrupt eyewitness identifications?. *Psychological Science, 20*, 122-126. doi: 10.1111%2Fj.1467-9280.2008.02262.x

Horry, R., Brewer, N., Weber, N., & Palmer, M. A. (2015). The effects of allowing a second sequential lineup lap on choosing and probative value. *Psychology, Public Policy, and Law, 21*, 121-133. doi: 10.1037/law0000041

Jalava, S. T., Smith, A.M., & Wells, G.L. (under review). The role of estimator variables in eyewitness identification: High confidence does not always imply high accuracy.

Johansson, P., Hall, L., Sikström, S., Tärning, B., & Lind, A. (2006). How something can be said about telling more than we can know: On choice blindness and introspection. *Consciousness and Cognition, 15*, 673-692. doi: 10.1016/j.concog.2006.09.004

Juslin, P., Olsson, N., & Winman, A. (1996). Calibration and diagnosticity of confidence in eyewitness identification: Comments on what can be inferred from the low confidence–accuracy correlation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 1304-1316. doi: 10.1037/0278-7393.22.5.1304

Karelitz, T. M., & Budescu, D. V. (2004). You say" probable" and I say" likely": improving interpersonal communication with verbal probability phrases. *Journal of Experimental Psychology: Applied, 10*, 25-41. doi: 10.1037/1076-898X.10.1.25

Kenchel, J., Reisberg, D., & Dodson, C. S. (2017). "*In your own words, how certain are you?" Post-identification feedback powerfully distorts verbal expressions of witness confidence.* Paper at the American Psychology-Law Society. Seattle, U.S.A.

Leippe, M.R., & Eisenstadt, D. (2007). In R.C.L. Lindsay, D.F. Ross, J.D. Read JD, & M.P. Toglia (Eds.). Eyewitness confidence and the confidence-accuracy relationship in memory for people. *The Handbook of Eyewitness Psychology, Volume II: Memory for People* (pp. 377–425). Mahwah: Lawrence Erlbaum Associates.

Lindsay, R. C., & Wells, G. L. (1985). Improving eyewitness identifications from lineups: Simultaneous versus sequential lineup presentation. *Journal of Applied Psychology, 70*, 556-564. doi: 10.1037/0021-9010.70.3.556

Luus, C. E., & Wells, G. L. (1991). Eyewitness identification and the selection of distracters for lineups. *Law and Human Behavior, 15*, 43-57. doi: 10.1007/BF01044829

Malpass, R. S., & Devine, P. G. (1981). Eyewitness identification: Lineup instructions and the absence of the offender. *Journal of Applied Psychology, 66*, 482-489. doi: 10.1037/0021-9010.66.4.482

Manson v. Braithwaite, 432 U.S. 98 (1977).

Mansour, J. K., Lindsay, R. C. L., Brewer, N., & Munhall, K. G. (2009). Characterizing visual behaviour in a lineup task. *Applied Cognitive Psychology, 23*, 1012-1026. doi: 10.1002/acp.1570

Mickes, L. (2015). Receiver operating characteristic analysis and confidence–accuracy characteristic analysis in investigations of system variables and estimator variables that affect eyewitness memory. *Journal of Applied Research in Memory and*

Cognition, 4, 93-102. doi: 10.1016/j.jarmac.2015.01.003

Murphy, A. H., Lichtenstein, S., Fischhoff, B., & Winkler, R. L. (1980). Misinterpretations of precipitation probability forecasts. *Bulletin of the American Meteorological Society, 61*, 695-701. doi: 10.1175/1520-0477(1980)061%3C0695:MOPPF%3E2.0.CO;2

National Research Council (2014). *Identifying the culprit: Assessing eyewitness identification.* Washington, DC: The National Academies Press. doi:10.17226/18891.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*, 231-259. doi: 10.1037/0033-295X.84.3.231

Oriet, C., & Fitzgerald, R. J. (2018). The single lineup paradigm: A new way to manipulate target presence in eyewitness identification experiments. *Law and Human Behavior, 42*, 1-12. doi: 10.1037/lhb0000272

Palmer, M. A., Brewer, N., Weber, N., & Nagesh, A. (2013). The confidence-accuracy relationship for eyewitness identification decisions: Effects of exposure duration, retention interval, and divided attention. *Journal of Experimental Psychology: Applied, 19*, 55-71. doi: 10.1037/a0031602

Police and Criminal Evidence Act 1984 (PACE) Code D Revised: Code of practice for the identification of persons by Police Officers. (February, 2017). London: Home Office. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/592562/pace-code-d-2017.pdf

R Core Team. (2019).

Rodriguez, D. N., & Berry, M. A. (2019). Administrator blindness affects the recording of eyewitness lineup outcomes. *Law and Human Behavior*. doi: 10.1037/lhb0000352

Sauer, J. D., Palmer, M. A., & Brewer, N. (2019). Pitfalls in using eyewitness confidence to

diagnose the accuracy of an individual identification decision. *Psychology, Public Policy, and Law*. doi:10.1037/law0000203

Semmler, C., Dunn, J., Mickes, L., & Wixted, J. T. (2018). The role of estimator variables in eyewitness identification. *Journal of Experimental Psychology: Applied, 24*, 400-415. doi: 10.1037/xap0000157

Smith, A. M., Wilford, M. M., Quigley-McBride, A., & Wells, G. L. (2019). Mistaken eyewitness identification rates increase when either witnessing or testing conditions get worse. *Law and Human Behavior, 43*, 358-368. doi: 10.1037/lhb0000334

Sporer, S. L. (1993). Eyewitness identification accuracy, confidence, and decision times in simultaneous and sequential lineups. *Journal of Applied Psychology, 78*, 22-33. doi: 10.1037/0021-9010.78.1.22

Sporer, S. L., Penrod, S., Read, D., & Cutler, B. (1995). Choosing, confidence, and accuracy: a meta-analysis of the confidence-accuracy relation in eyewitness identification studies. *Psychological Bulletin, 118*, 315-327. doi: 10.1037/0033-2909.118.3.315

Steblay, N. K. (2011). What we know now: The Evanston Illinois field lineups. *Law and Human Behavior, 35*, 1-12. doi: 10.1007/s10979-009-9207-7

Tryon, W. W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: an integrated alternative method of conducting null hypothesis statistical tests. *Psychological Methods, 6*, 371-386. doi: 10.1037/1082-989X.6.4.371

Vredeveldt, A., & Sauer, J. D. (2015). Effects of eye-closure on confidence-accuracy relations in eyewitness testimony. *Journal of Applied Research in Memory and Cognition, 4*, 51-58. doi: 10.1016/j.jarmac.2014.12.006

Wallsten, T. S., & Budescu, D. V. (1995). A review of human linguistic probability processing: General principles and empirical evidence. *The Knowledge Engineering*

Review, 10, 43-62. doi: 10.1017/S0269888900007256

Wallsten, T. S., Budescu, D. V., Rapoport, A., Zwick, R., & Forsyth, B. (1986). Measuring the vague meanings of probability terms. *Journal of Experimental Psychology: General, 115*, 348-365. doi: 10.1037/0096-3445.115.4.348.

Wallsten, T. S., Budescu, D. V., & Zwick, R. (1993). Comparing the calibration and coherence of numerical and verbal probability judgments. *Management Science, 39*, 176-190. doi: 10.1287/mnsc.39.2.176

Wallsten, T. S., Fillenbaum, S., & Cox, J. A. (1986). Base rate effects on the interpretations of probability and frequency expressions. *Journal of Memory and Language, 25*, 571-587. doi: 10.1016/0749-596X(86)90012-4

Weber, E. U., & Hilton, D. J. (1990). Contextual effects in the interpretations of probability words: Perceived base rate and severity of events. *Journal of Experimental Psychology: Human Perception and Performance, 16*, 781-789. doi: 10.1037/0096-1523.16.4.781

Weber, N., Brewer, N., & Margitich, S. (2008). The confidence-accuracy relation in eyewitness identification: Effects of verbal versus numeric confidence scales. In K. H. Keifer (Ed.), *Applied Psychology Research Trends* (pp. 103-118). New York, NY: Nova Publishers.

Wells, G. L., & Bradfield, A. L. (1998). "Good, you identified the suspect": Feedback to eyewitnesses distorts their reports of the witnessing experience. *Journal of Applied Psychology, 83*, 360-376. doi: 10.1037/0021-9010.83.3.360

Wicklin, R. (2017, July 12). The bias-corrected and accelerated (BCa) bootstrap interval [blog]. Retrieved from https://blogs.sas.com/content/iml/2017/07/12/bootstrap-bca-interval.html

Windschitl, P. D., & Wells, G. L. (1996). Measuring psychological uncertainty: Verbal

versus numeric methods. *Journal of Experimental Psychology: Applied*, *2*, 343-364. doi: 10.1037/1076-898X.2.4.343 .

Wixted, J. T., Mickes, L., Clark, S. E., Gronlund, S. D., & Roediger III, H. L. (2015). Initial eyewitness confidence reliably predicts eyewitness identification accuracy. *American Psychologist, 7*0, 515-526. doi: 10.1037/a0039510.

Wixted, J. T., Mickes, L., Dunn, J. C., Clark, S. E., & Wells, W. (2016). Estimating the reliability of eyewitness identifications from police lineups. *Proceedings of the National Academy of Sciences, 113*, 304-309. doi: 10.1073/pnas.1516814112

Wixted, J. T., Read, J. D., & Lindsay, D. S. (2016). The effect of retention interval on the eyewitness identification confidence–accuracy relationship. *Journal of Applied Research in Memory and Cognition, 5*, 192-203. doi: 10.1016/j.jarmac.2016.04.006

Wixted, J. T., & Wells, G. L. (2017). The relationship between eyewitness confidence and identification accuracy: A new synthesis. *Psychological Science in the Public Interest, 18*, 10-65. doi: 10.1177/1529100616686966

Yaniv, I., Yates, J. F., & Smith, J. K. (1991). Measures of discrimination skill in probabilistic judgment. *Psychological Bulletin, 110*, 611-617. doi: 0.1037/0033-2909.110.3.611

Table 1

Proportion Identification Responses as a Function of Target for Experiments 1 and 2

Target	Response		
	Suspect ID	Filler ID	Rejection
	Experiment 1		
Target-present lineups			
164[a]	.31	.42	.26
170[b]	.60	.10	.30
173 [a]	.71	.11	.18
176 [b]	.38	.32	.30
Target-absent lineups			
164[a]	.07	.30	.63
170[b]	.09	.14	.76
173 [a]	.04	.36	.60
176 [b]	.03	.21	.75
	Experiment 2		
Target-present lineups			
114[c]	.61	.24	.15
125[d]	.64	.12	.24
170 [d]	.70	.14	.15
173[c]	.59	.28	.13
Target-absent lineups			
114[c]	.09	.42	.50
125[d]	.14	.16	.70
170 [d]	.16	.28	.56
173[c]	.09	.32	.58

Note: Targets sharing a subscript were yoked together in an experiment such that the target-present lineup for one served as the target-absent lineup for the other.

Table 2

Calibration Statistics for each Condition in Experiments 1 and 2

Condition	C		O/U		ANDI	
Experiment 1						
scale only	.03 (.02)	[.003, .060]	0.01 (0.05)	[-0.08, 0.10]	.13 (.08)	[.000, .240]
own words	.01 (.01)	[.001, .023]	0.10 (0.05)	[0.01, 0.19]	.15 (.07)	[.010, .283]
why	.09 (.03)	[.039, .142]	0.12 (0.06)	[0.01, 0.23]	.02 (.05)	[.000, .098]
selection	.04 (.02)	[.012, .070]	0.16 (0.05)	[0.07, 0.26]	.12 (.08)	[.000, .240]
scale rating second	.03 (.01)	[.013, .048]	0.12 (0.03)	[0.07, 0.18]	.07 (.03)	[.008, .125]
overall	.03 (.01)	[.011, .041]	0.10 (0.02)	[0.05, 0.14]	.08 (.03)	[.024, .132]
Experiment 2						
scale only	.02 (.01)	[.004, .029]	0.10 (0.03)	[0.04, 0.16]	.11 (.04)	[.025, .185]
own words	.03 (.01)	[.012, .052]	0.15 (0.03)	[0.08, 0.21]	.06 (.04)	[.003, .123]
overall	.02 (.01)	[.011, .034]	0.12 (0.02)	[0.08, 0.17]	.09 (.03)	[.033, .137]

Note: Parentheses contain standard deviations calculated using 10,000 bootstrapped samples. Confidence intervals are 95% BCa confidence intervals.

Table 3

Inferential Confidence Intervals for Comparisons Between Confidence Conditions

Condition	C		O/U		ANDI	
Experiment 1						
scale only vs. own words	[.000, .060]	[.000, .032]	[-0.070, 0.091]	[0.020, 0.175]	[.000, .254]	[.019, .286]
scale only vs. why	[.000, .060]	[.039, .139]	[-0.071, 0.091]	[0.025, 0.214]	[.000, .256]	[.000, .114]
scale only vs. selection	[.000, .059]	[.007, .071]	[-0.070, 0.091]	[0.078, 0.242]	[.000, .254]	[.003, .246]
own words vs. why	[.000, .033]	[.036, .142]*	[0.020, 0.175]	[0.025, 0.214]	[.017, .288]	[.000, .114]
own words vs. selection	[.000, .032]	[.006, .071]	[0.020, 0.175]	[0.078, 0.242]	[.019, .286]	[.003, .246]
why vs. selection	[.039, .139]	[.006, .071]	[0.025, 0.214]	[0.078, 0.242]	[.000, .113]	[.002, .248]
scale only vs. all other conditions	[.000, .054]	[.015, .044]	[-0.057, 0.078]	[0.083, 0.166]*	[.013, .238]	[.016, .116]
Experiment 2						
scale only vs. own words	[.005, .028]	[.015, .048]	[0.056, 0.141]	[0.10, 0.19]	[.046, .170]	[.008, .115]

Note: Inferential confidence intervals (ICIs) are 95% BCa confidence intervals. ICIs are ordered in line with the order that conditions are listed. For example, for the scale only vs. own words comparison, the ICI for Scale only is presented on the left and the ICI for own words on the right. A Bonferroni correction was used for the Experiment 1 comparisons of individual conditions. * indicates a significant difference between pairs of ICIs.

Table 4

Correspondence Between Scale Ratings and Verbal Confidence Statements for Identifications in Experiments 1 and 2

	Scale Rating		
Verbal Confidence Condition	Low	Medium	High
	Experiment 1		
Own words			
Low (*n* = 24)	.79	.17	.04
Medium (*n* = 55)	.09	.75	.16
High (*n* = 46)	.00	.06	.94
Why			
Low (*n* = 16)	.25	.38	.38
Medium (*n* = 92)	.10	.44	.46
High (*n* = 7)	.00	.00	1.00
Uninterpretable (*n* = 2)	.50	.00	.50
Selection			
Low (*n* = 27)	.48	.44	.07
Medium (*n* = 60)	.03	.53	.43
High (*n* = 25)	.04	.04	.92
	Experiment 2		
Own words			
Low (*n* = 26)	.69	.27	.04
Medium (*n* = 135)	.03	.63	.34
High (*n* = 97)	.01	.04	.95
Uninterpretable (*n* = 8)	.12	.12	.75

Note: This table include all identifications from target-present and target-absent lineups (including filler identifications from target-present lineups).

Supplemental Material - Calculation and Comparison of Calibration Statistics

Following Vredeveldt and Sauer (2015) and based on the advice of Yaniv, Yates, and Smith (1991), the adjusted normalized discrimination index (ANDI) was calculated rather than the adjusted normalized resolution index (ANRI) as preregistered. Following Palmer, Brewer, Weber, and Nagesh (2013) and the process suggested by Tryon (2001), inferential confidence intervals (ICI) were calculated for calibrations statistics for comparisons. All calculated confidence intervals were bias-corrected and accelerated (BCa) confidence intervals. BCa confidence intervals are adjusted for skewness in the bootstrap distribution used to calculate the standard deviation for each parameter and based on the original data rather than the bootstrapped data as in the case for standard confidence intervals (DiCiccio & Efron, 1996; Wicklin, 2017). All bootstrapping was conducted in R (Canty & Ripley, 2019; R Core Team, 2019)

Supplemental Material - Supplemental Analyses

Experiment 1

Comparing to real eyewitnesses. As the participants in the current study experienced a very different situation than real eyewitness, I examined whether participants performed similarly to real eyewitnesses. Supplemental Table 3 provides the frequency with which the statements similar to those listed by Behrman and Richards (2005) were given by choosers in the own words, why, and selection conditions; the table also indicates how often novel statements were obtained from participants. The statements recorded by Behrman and Richards from real eyewitnesses were commonly reported by the current participants although the language used often differed slightly. Most of these aligned closely in meaning to the statements Behrman and Richards obtained from real participants. For example, a frequently made statement in this experiment was “I’m not very confident” which although not reported by Behrman and Richards, corresponds closely to their statement “I think, but I

am not sure, I made the right decision".

I examined whether confidence as categorized based on Behrman and Richards' (2005) scheme was similarly predictive of decisions in the current sample as in their sample. For this analysis, suspect identifications comprised target identifications from target-present lineups and identifications of the designated innocent suspect in target-absent lineups; all other lineup identifications were treated as filler identifications, including those from target-present lineups. Behrman and Richards found a ratio of approximately 3:1 for suspect compared to filler identifications; in the current sample, however, the ratio was approximately 1:1. Logistic regression was used to determine whether confidence query (scale only, own words, why, selection), confidence level (low, medium, high), and/or their interaction predicted whether an identification was of a suspect or filler, regardless of accuracy. The model was significant, χ^2 (7, n = 475) = 31.28, $p < .001$, $r_{\text{Nagelkerke}} = .08$. Confidence level was the only significant predictor, $\chi^2(1) = 22.50$, $p < .001$, $OR = 2.24$. The model improved categorization of cases from 52.4% to 60.4%. The overall result then is consistent with Behrman and Richards; however, they reported a point biserial correlation coefficient for their effect size rather than Nagelkerke's *r*. The point biserial correlation coefficient for the current sample, $r(473) = .29$, $p < .001$, is similar, though somewhat higher, than that reported by Behrman and Richards ($r = .21$). An exploratory z-test indicated the difference was not significant ($p = .31$).

Following a verbal judgment of confidence with a scale rating. In order to determine whether scale ratings of confidence might be different if asked for after eyewitnesses assessed their confidence using a verbal means versus if participants were only asked for a scale rating, I conducted a univariate analysis of variance (ANOVA) with confidence query (scale only, own words, selection, why) and choosing (identification, rejection) as the predictors and the scale rating as the measure. Only the interaction was

significant, $F(3, 904) = 3.82$, $p = .010$, $\eta_p^2 = .01$ (all other ps > .64). There was no significant effect of confidence query for rejections (all pairwise ps > .069). For identifications, the scale only condition resulted in lower confidence ($M = 63.84$, $SD = 25.79$) than in the selection ($M = 72.15$, $SD = 23.53$; $p = .010$) and why ($M = 72.07$, $SD = 21.79$; $p = .010$) conditions but did not differ from the own words condition ($M = 66.89$, $SD = 23.21$; $p = .33$). The relevant CAC curves are presented in Supplemental Figure 1.

Yoking analysis. An exploratory analysis was conducted to evaluate the predictive ability of scale only compared to own words judgments of confidence. Following Kenchel, et al. (2017), I treated the numeric confidence rating provided by participants after their verbal confidence rating as a translation of their verbal confidence rating into a numeric value. I then yoked participants who made scale only judgements to participants who made verbal statements translated with that same numeric value. This coding resulted in 161 pairs of participants ($n = 322$). A logistic regression with confidence query (scale only, own words), confidence level (0-100), and their interaction was conducted to predict the accuracy of identifications and then the accuracy of rejections. For identifications, the model was significant, $\chi^2(3) = 29.74$, $p < .001$, $r_{Nagelkerke} = .15$, but only confidence level predicted accuracy, $\chi^2(1) = 9.07$, $p = .003$, $OR = 1.02$ (all other ps > .21). Classification accuracy improved from 52.8% to 64.5% with the predictors. For rejections, the model was not significant, $\chi^2(3) = 6.18$, $p = .10$, $r_{Nagelkerke} = .04$, although confidence was a significant predictor of accuracy, $\chi^2(1) = 4.32$, $p = .038$, $OR = 1.02$ (all other ps > .18). Classification accuracy did not improve with inclusion of the predictors (75.7%). Thus, as expected, own words and scale judgments of confidence did not differ in their predictive ability.

Experiment 2.

Comparing to real eyewitnesses. Supplemental Table 3 illustrates the extent to which statements obtained in the own words condition corresponded with the statements

reported by Behrman and Richards (2005). A logistic regression with confidence query (scale only, own words), confidence level (low, medium, high), and their interaction were entered as predictors of identification accuracy. The model was significant, χ^2 (3, n = 527) = 24.94, p < .001, $r_{\text{Nagelkerke}}$ = .06. Level of confidence was the only significant predictor, $\chi^2(1) = 9.58$, p = .002, OR = 1.72 (all other ps > .27). Categorization improved from 61.1% with the no predictor model to 62.6% with the predictors added. Similar to Behrman and Richards and Experiment 1, the CA correlation was moderate, $r(525) = .27$, $p < .001$.

Following a verbal judgment of confidence with a scale rating. The next analysis assessed whether scale-based confidence ratings differed as a function of whether they were obtained on their own (scale only condition) versus after the provision of a verbal judgment (own words condition). An ANOVA with confidence query and choosing (identification, rejection) was conducted on the scale confidence ratings. As in Experiment 1, only the interaction was significant, $F(3, 855) = 4.42$, $p = .036$, $\eta_p^2 = .005$ (all other ps > .13). There was no effect of confidence query for rejections (p = .26), but there was an effect on identifications (p = .050). For identifications, scale ratings in the scale only condition (M = 70.38, SD = 24.20) were lower than in the own words condition (M = 74.78, SD = 22.82), $t(534) = 2.16$, $p = .031$, $d = 0.19$.[8]

Yoking analysis. As in Experiment 1, I conducted an exploratory analysis wherein I yoked participants in the own words and scale only conditions on basis of their scale ratings. Doing so resulted in 336 pairs. A logistic regression on identifications was conducted with confidence query, confidence level (0-100), and their interaction as predictors. The model was significant, $\chi^2(3, n = 410) = 39.81$, $p < .001$, $r_{\text{Nagelkerke}} = .12$, however, confidence level was the only significant predictor, $\chi^2(1) = 22.90$, $p < .001$, OR = 2.48 (all other ps > .11).

[8] The interaction was marginally significant if target-present filler identifications were excluded from analysis (p = .056 for the interaction, p = .11 for the simple effect of condition on identifications).

Including the predictors improved classification accuracy from 52.9% to 63.4%.

When this analysis was replicated with only suspect identifications, the model was again significant, $\chi^2(3, n = 248) = 26.60$, $p < .001$, $r_{Nagelkerke} = .19$, but both confidence level, $\chi^2(1) = 16.69$, $p < .001$, $OR = 1.06$, and confidence query, $\chi^2(1) = 5.24$, $p = .022$, $OR = 18.94$, were significant. Participants were 18.94 times more likely to make a correct identification in the own words condition than the scale only condition. The interaction was marginally significant, $\chi^2(1) = 3.33$, $p = .068$, $OR = 0.97$, suggesting that confidence was a slightly better predictor of accuracy in the scale only condition than the own words condition. However, this analysis must be interpreted with caution as the number of innocent suspect identifications was considerably lower than the number of correct identifications; there was no change in the accuracy of classification when the predictors were included compared to when the model contained no predictors (87.5%).

Finally, rejection accuracy was considered using the same model predictors, $\chi^2(3, n = 261) = 21.90$, $p < .001$, $r_{Nagelkerke} = .12$. Only confidence level was significant, $\chi^2(1) = 12.63$, $p < .001$, $OR = 1.03$ (all other ps $> .25$). Classification accuracy improved from 77.0% to 77.4% in the full model.

Supplemental Table 1:

Options Participants Could Choose from to Indicate their Confidence in their Lineup Decision in the Selection Condition of Experiment 1

Statement	Confidence Rating
The person I chose looked kind of like him.	Low
The person I chose looked similar to him	
The person I chose looked sort of like him.	
The person I chose looked familiar*	
I am not quite sure about my decision.	
I am not exactly certain about my decision.	
The person I chose looks somewhat like him.	
The person I chose resembles him.	
The person I chose may have been him.	
The person I chose could be him.	
The person I chose is the closest to him.	
The person I chose is possibly him.	
I think, but I am not sure, I made the right decision.	
The person I chose looked very similar to him.	Medium
The person I chose looks most like him.	
I am pretty sure I made the right decision.	
The person I chose looks like him.	
I am almost certain I made the right decision.	
The person I chose is almost a perfect match to him.	
I am not 100% certain of my decision.	
I think the person I chose did it.	
I believe the person I chose did it.	
I am fairly certain about my decision.	
I am relatively sure about my decision.	
I am moderately sure about my decision.	
I am absolutely certain about my decision.	High
I will never forget his face.	
I am certain the person I chose did it.	
I am positive the person I chose did it.	
That's him, I don't need to see any other photos.	
I would testify in court that it is him.	
I am sure that it is him.	
It is definitely him.	
I am very sure about my decision.	
The person I chose looks exactly like him.	

Note: The assigned confidence rating corresponded with the ratings assigned by Behrman and Richards (2005) for almost identical statements. The item denoted with a * was accidentally not included as an option in the selection condition even though a similar statement appeared in Behrman and Richards.

Supplemental Table 2

Statements by Identifiers in Experiments 1 and 2 that did not Correspond with Statements Used in Behrman and Richards (2005)

Own Words Condition	Rating		Confidence
	M	*SD*	Category
Not confident at all.	0.97	1.80	Low
Not confident.	1.19	1.77	Low
Not very confident.	2.14	1.93	Low
I do not feel very confident.	2.22	2.00	Low
I am not very confident.	2.25	1.96	Low
Not much.	2.28	1.85	Low
I would have to see the incident again.	2.69	3.17	Low
I'm not real confident. He looks similar.	3.06	1.87	Low
I give my best guess.	3.17	2.76	Low
Not too confident as I saw most of a profile in the video and not the face from the front.	3.25	1.66	Low
I am not sure if I selected the right person. They all have similar characteristics, so I can't say with 100% certainty that I selected the right person.	3.28	2.09	Low
Partially .	3.44	1.76	Low
Slightly confident.	3.67	2.16	Low
I'm not that confident, only saw him from the side but I think that's the guy.	3.69	2.07	Low
Not 100%.	4.42	2.78	Medium*
I think number six ins the one not 100% positive	4.50	2.65	Medium
I'm somewhat confident.	4.50	1.99	Medium
I am slightly confident.	4.53	1.96	Medium
I think so because he has the same nonchalant face.	4.58	1.87	Medium
I am not extremely confident that number 5 is the criminal. However, I will say that if the criminal is definitely in the line up, I am quite confident that it is number 5.	4.64	2.45	Medium
Good.	5.00	2.15	Medium
I feel I am somewhat confident because of his features.	5.19	2.10	Medium
I am reasonably certain that I have the right guy but not positive.	5.25	2.02	Medium
His eye brows were very curvy I noticed when he was stealing the money. Although he was not smiling while stealing and he is smiling in the lineup, his eye brows give him away.	5.56	2.83	Medium
I have to say that I am pretty clear about my choice, the only thing throwing me off right now is the fact that he may not be there. There is only one choice I would make of the six though so I am trusting my hunch.	5.61	2.45	Medium
I am very confident since I watched the video twice. however I did not see a frontal shot of the guy only a side of his face.	5.64	2.44	Medium
Yes confident.	7.22	2.33	Medium
I am very confident that that is the person.	7.25	2.84	High*
Very accurate.	7.42	2.60	High*
I am completely confident in my choice.	8.61	1.95	High
Why Condition			
No.	2.39	3.49	Low
I picked the second one .	2.75	2.30	Low
Number 1 looked like he could be the same person, maybe with a haircut.	3.14	2.42	Low
Not sure but looks same person.	3.47	2.20	Low
His attitude	3.56	2.40	Low
The culprit it like him.	3.81	2.34	Low
It was the person closet to the person in the video.	4.08	2.57	Medium*
Looks.	4.28	2.54	Low
I think # 6 looked like the person who took something out of the purse. It was not his expression that convinced me.	4.39	2.27	Medium*

He looked the most like the criminal.	4.44	2.56	Medium*
Person had curly black hair.	4.61	2.80	Medium
Looks right.	4.64	2.80	Medium
It was number three if any of them. No one else looked like the person who stole the money.	4.67	2.95	Medium
I noticed him.	4.75	2.63	Medium
His features and hair color and style are similar.	4.81	2.25	Medium
The jaw line.	4.83	2.25	Medium
Similar hairstyle and facial structure.	4.86	2.40	Medium
To the best of my memory, this was the boy who committed the crime.	4.86	2.34	Medium
I CHOSE THAT PERSON BECAUSE IT LOOKED LIKE THE MOST SIMILAR TO THE.CRIMINAL	4.89	2.88	Medium
This person is the closest to the one I saw in the video	5.06	2.48	Low*
Matches the look.	5.08	2.67	Medium
His skin complexation and hair seems to fit the image of the thief. His eyes also looked familiar.	5.11	2.58	Medium
The nose was distinguishable.	5.11	2.56	Medium
His skin complexation and hair seems to fit the image of the thief. His eyes also looked familiar.	5.11	2.58	Medium
I recall the person having dark hair and rather bushy eyebrows.	5.14	2.31	Medium
Their brow line looked distinctive like the person in the video.	5.17	2.48	Medium
They person from the video and lineup have similar facial features and hairstyles.	5.17	1.81	Medium
The person I chose, stood out from all the other images.	5.19	1.94	Medium
I chose the person because he looked like the same guy on the video stealing the money.	5.25	2.45	Medium
Looks like same hair color, eyes and face.	5.25	2.57	Medium
The hair and face look the same.	5.25	2.48	Medium
I believe number 1 is our suspect.	5.28	2.87	Medium
The facial features looked to be the man in the video.	5.31	2.62	Medium
It was the guy officer.	5.33	2.91	Medium
The face looks similar to the person who did the crime, especially the hair.	5.33	2.54	Medium
This person most resembles the person I saw in the video.	5.33	2.53	Medium
I believe this was the man from the crime scene.	5.42	2.45	Medium
The facial features match the guy in the office.	5.47	2.63	Medium
The length of hair and facial features matched.	5.47	2.42	Medium
The person accused is in the line up he had brown hair, brown eyes, and that same profile it fits.	5.50	2.72	Medium
His face is the same.	5.58	2.85	Medium
The hair color and length, also facial features.	5.67	2.37	Medium
It looked just like him I remember his slender face.	5.75	2.56	Medium
It looks like him so much! I remember his eyes and eyebrows, but wish i could see a side view.	5.75	2.16	Medium
I truly think that is the boy in the video, he had a certain "look".	5.86	3.18	Medium
Number 4 looked very much like the person I saw in the video. I am sure it is the same person.	5.89	2.74	Medium
Recognized the face from the video.	6.06	2.94	Medium
This person looks just like the man in the video that stole the money from the purse.	6.14	2.63	Medium
I am really good at identifying facial bone structure. That guy's bone structure looks similar to the one I saw in the video.	6.19	2.91	Medium
Is him.	6.22	3.48	Medium
I picked him because I memorized the look of his face and hair.	6.25	3.03	Medium
I remembered his face.	6.31	2.86	Medium
That is the face of the person I remember seeing who stole the money.	6.33	2.99	Medium
That one.	6.39	3.33	Medium
I remember him.	6.42	2.71	Medium
That was the guy that I saw, I recognized that haircut anywhere.	6.42	2.84	Medium
I chose 6 because that was the guy.	6.56	3.20	Medium
I know that it is number 3	6.69	3.52	Medium

That is the person who I saw take the money out of the bag.	6.69	3.13	Medium
The person was the one I saw in the video.	6.72	3.12	Medium
Number 5 is definitely the person from the video. I have no doubts at all.	8.06	3.18	High

Note: Additional statements were collected; however, the set was pared down to remove redundancy as often multiple participants made similar statements. Statements were rated on a scale of 0 (No Confidence) to 10 (Absolute Certainty) and statements were assigned to confidence categories following Behrman & Richards (2005) as much as possible; average ratings were rounded to their nearest integer to determine category placement. *indicates statements whose mean ratings did not align with Behrman and Richards' coding scheme and so these ratings were not used to determine category placement.

Supplemental Table 3

Frequency of Behrman and Richards' (2005) Confidence Statements by Identifiers in Experiment 1 and 2

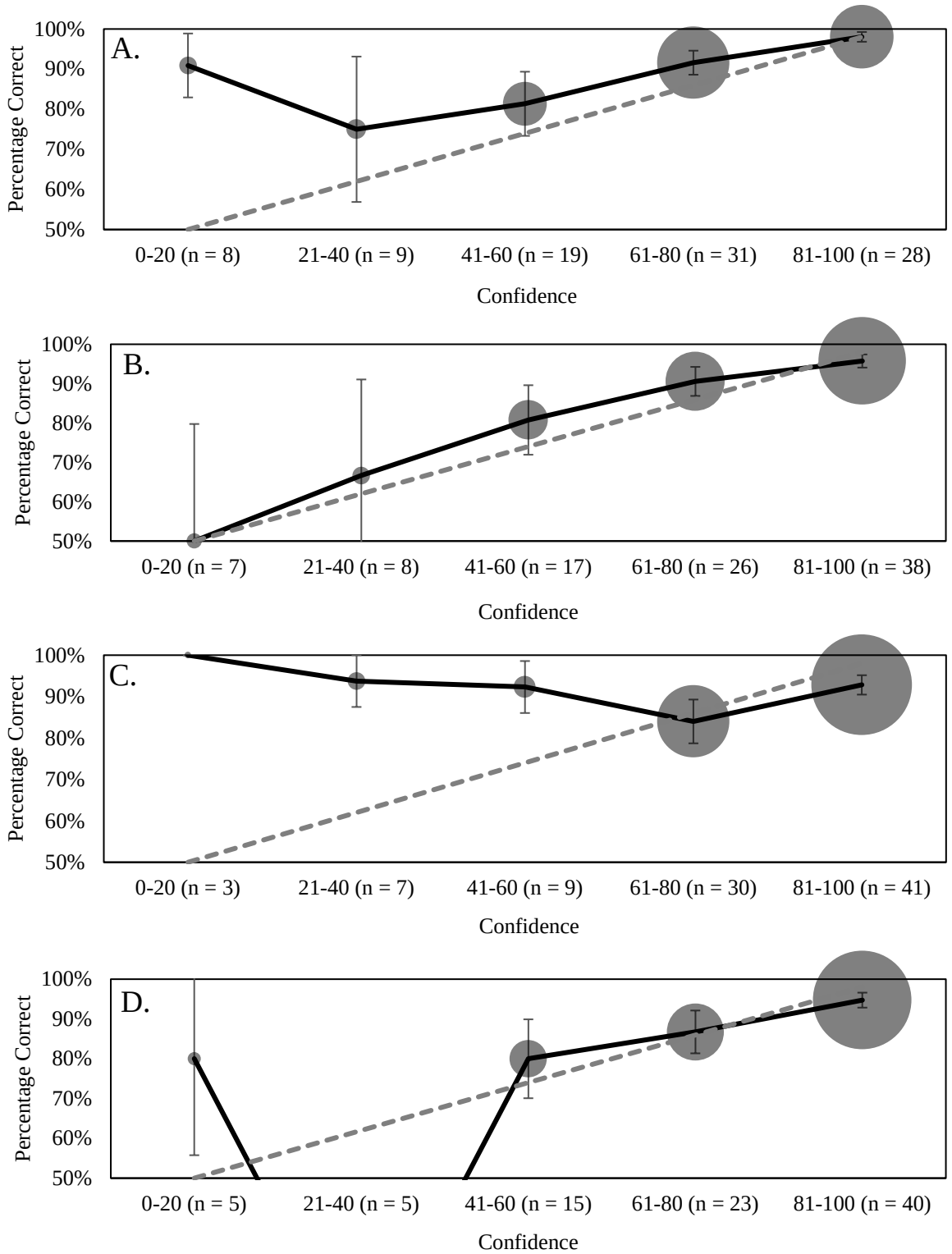
Statement	Condition			
	Selection (n = 112)	Own Words – Exp. 1 (n = 125)	Why (n = 117)	Own Words – Exp. 2 (n = 267)
0-4 Confidence Level				
The person I chose looked kind of like him.	7	0	0	0
The person I chose looked similar to him.	1	0	5	0
The person I chose looked sort of like him.	1	0	0	0
I am not quite sure about my decision.	4	0	0	4
I am not exactly certain about my decision.	2	0	0	6
The person I chose looks somewhat like him.	1	0	0	0
The person I chose resembles him.	0	0	6	1
The person I chose may have been him.	1	0	0	0
The person I chose could be him.	0	0	1	0
The person I chose is the closest to him.	0	0	7	0
The person I chose is possibly him.	6	0	0	0
I think, but I am not sure, I made the right decision.	4	0	0	2
5-7 Confidence Level				
The person I chose looked very similar to him.	9	0	4	1
The person I chose looks most like him.	3	0	9	0
I am pretty sure I made the right decision.	12	6	0	26
The person I chose looks like him.	5	1	30	2
I am almost certain I made the right decision.	7	0	0	2
The person I chose is almost a perfect match to him.	2	0	1	2
I am not 100% certain of my decision.	4	5	0	5
I think the person I chose did it.	3	0	7	1
I believe the person I chose did it.	3	0	3	0
I am fairly certain about my decision.	3	11	0	13
I am relatively sure about my decision.	0	0	0	0
I am moderately sure about my decision.	9	1	0	12
8-10 Confidence Level				
I am absolutely certain about my decision.	9	1	1	0
I will never forget his face.	1	0	0	1
I am certain the person I chose did it.	0	2	0	6
I am positive the person I chose did it.	1	0	0	1
That's him, I don't need to see any other photos.	2	0	0	1
I would testify in court that it is him.	1	0	0	0
I am sure it is him.	3	2	1	1
It is definitely him.	2	0	0	0
I am very sure about my decision.	4	19	0	23
The person I chose looks exactly like him.	2	1	3	4
Not Captured by Behrman and Davey (2005)				
Numeric rating (with or without additional text)	-	46	5	119
Statement related to confidence	-	29	31	17
Statement unrelated to confidence	-	0	6	6

Note: the words *accurate*, *sure*, *certain*, and *confident* were treated as interchangeable.

Supplemental Table 4

Number of Lineup Laps and Frequency of Multiple Selections in Experiment 2

Laps of Lineup	Lineup members selected							
	0	1	2	3	4	5	6	Total
Final Decision was to Identify a Lineup Member								
1	0	333	48	3	0	0	1	385
2	0	120	19	2	0	0	0	142
3	0	7	1	1	1	0	0	9
Overall	0	460	68	6	1	0	1	536
Final Decision was to Reject the Lineup								
1	208	0	1	0	0	0	0	209
2	111	0	0	0	0	0	0	111
3	2	0	0	1	0	0	0	3
Overall	321	0	1	1	0	0	0	323
Overall Total	321	460	69	7	1	0	1	859



Supplemental Figure 1. Confidence-Accuracy Characteristic (CAC) curves for the scale ratings in Experiment 1 for participants who provided A) only a scale rating or a scale rating after providing B) their confidence in their own words, C) explaining why they chose the person they did, or D) indicating their confidence by selecting from a series of statements. The size of the plotted data point reflects the proportion of cases that contributed to the calculation (as suggested by Evelo, et al., 2018). Standard error bars were estimated using 10,000 bootstrapped samples.